Identifying genetic susceptibilities underlying familial haematological malignancies in a Tasmanian family resource

by

Nicholas Bayden Blackburn, BSc (Hons)

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy

University of Tasmania

August 2015

Declaration of Originality

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of the my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Nicholas Bayden Blackburn

Authority of Access Statement

This thesis may be made available for loan and limited copying and communication in accordance with the *Copyright Act 1968*.

Nicholas Bayden Blackburn

Statement of Ethical Conduct

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University.

Nicholas Bayden Blackburn

Statement of Co-Authorship

Nicholas B. Blackburn has incorporated a version of the published open access paper Blackburn *et al*.[1] into Chapter 5 of his dissertation. The article is included in Appendix 5.1.

This work was co-authored with: Dr Jac C. Charlesworth, Mr James R. Marthick, Dr Elizabeth M. Tegg, Dr. Katherine A. Marsden, A/Prof Velandai Srikanth, Prof John Blangero, Prof Ray M. Lowenthal, Prof Simon J. Foote and A/Prof Joanne L. Dickinson, with the majority of the research, analyses, text and figures generated and prepared by Nicholas himself.

Blackburn, N. B. *et al.* A retrospective examination of mean relative telomere length in the Tasmanian Familial Hematological Malignancies Study. *Oncology Reports* **33,** 25–32 (2015).

The contribution of the authors to the publication is as follows:
N.B.B. was the primary author, conducted data analysis, was involved in laboratory experiments and study design and wrote the manuscript; J.C.C. (co-first author) design and performance of statistical analyses, manuscript preparation; J.R.M. design and performance of telomere length assay, data analysis and DNA extractions; V.S. scientific input and permission for use of control samples; J.B. scientific advice and design of statistical methods; K.A.M., R.M.L. provided clinical expertise. E.M.T. provided clinical expertise and scientific input. S.J.F. assisted with overall scientific direction and provided input into study design. J.L.D. conceived and designed experiments and contributed to data analysis and manuscript preparation; and all authors contributed to the final manuscript preparation.

Signed: _____ _____

A/Prof Joanne Dickinson
Supervisor
Menzies Institute for Medical Research
University of Tasmania

Prof Tom Marwick
Director
Menzies Institute for Medical Research
University of Tasmania

Date: 11 August 2015 _____ 11 August 2015 _____

Acknowledgements

When I, rather naïvely, began this project four years ago I could never have imagined the opportunities that would grow from it, both personally and professionally. I am indebted to my supervisors, A/Prof Joanne Dickinson and Dr Jac Charlesworth, who together, both in complement and independently, have been amazing role models for me. Under their mentorship my fascination for genetics, and the dark-arts of bioinformatics, was encouraged and given the room it needed to grow. Thank you for all that you have taught me and for supporting me unwaveringly through various life and PhD-life challenges that have arisen. Thank you also to Prof Simon Foote for your initial supervision and involvement in this project.

I'm very grateful to my fellow Team-Y member (and Captain), Mr James Marthick, for your guidance, hands-on support and humour in relation to all lab related activities. I like to think we successfully proved in the early days that the combination of two Ys was best for southern blotting – others will of course beg to differ. Thank you especially for your work and dedication to establishing the telomere length assay which was pivotal to a body of work that saw me through many lab meetings, conferences and talks.

Mrs Annette Banks, your knowledge of all things related to Tasmanian family genealogy is second to none. Thank you for your genealogical support, patience in handling my oft complex queries about specific families and for being a smiling and encouraging face that has greeted me on many mornings (even if I did bribe you with lollies at various points in time).

A warm 'thank you' to the other members of the Cancer, Genetics and Gene Regulation Groups, both past and present, particularly Dr Kate Brettingham-Moore for your frequent coffee excursions, exchange of ideas, celebrations and woes. As well as A/Prof Kathryn Burdon for bringing fresh light to many aspects of my research and Dr Joy Rathjen for guidance when I needed it and tea-related sympathies. And a special note of gratitude to my fellow PhD students and colleagues, Naseem Ali, David Ward, Marie Buscot, Emma Cazaly and Ciarán O'Mara, for your friendship and support during both the fun and the difficult times. My friendships with each of you have been one of the major highlights of this PhD. Naseem and David, I look forward to wearing the appropriate attire with you when we are indoctrinated together.

I owe much to my Texan based colleagues. Particularly to Prof John Blangero for inviting me to train in the intricacies of NGS in his laboratory. Dr Joanne Curran, Dr Matthew Johnson and Dr Tom Dyer for your support, friendship and hospitality before, during and after my visit. And especially to Mr Juan Peralta – without your generous support, patience, friendship, extreme bioinformatics skills and instructions to 'gts', I'd still be finger painting. Thank you for being an ear of support and sympathy for many frustrations, sinceramente gracias amigo.

A personal thank you to my friends and family. Particularly Gustaf, Ella and Jess, you have each been a source of strength in each challenge that has arisen in the past four years and sounding boards for so many problems. Finally, Mum and Dad, thank you

for your constant support across four years of varying levels of grumpiness. Without you this would not have been possible. It is to each of you I dedicate this work.

More specifically, particular acknowledgement needs to be made to the following people who have contributed to this project.

The TFHMS clinicians, Prof Ray Lowenthal, Dr Katherine Marsden and Dr Elizabeth Tegg, for their assistance and input with the interpretation of the clinical aspects of, and records of patients in, this project.

The TFHMS genealogist, Mrs Annette Banks for her ongoing work with this familial resource.

Research Assistant James Marthick for his assistance across the laboratory aspects of this project.

A/Prof Kathryn Burdon and Sionne Lucas for their contributions to the amplicon based sequencing experiment.

The Australian Antarctic Division for access to and use of their ABI 3100 sequencer.

Mr Alistair Chilcott for IT support.

A/Prof Velandai Srikanth for his contributions to the telomere length work.

Prof John Blangero, Dr Tom Dyer and Mr Juan Peralta for their bioinformatics advice and support with this project

Prof Simon Foote for his initial involvement and supervision in this project.

The Australian Government for my Australian Postgraduate Award scholarship, and the University of Tasmania and Menzies Institute for Medical Research for additional scholarship and travel funding.

The participants of the Tasmanian Familial Haematological Malignancies Study.

Finally, A/Prof Joanne Dickinson and Dr Jac Charlesworth for their supervision throughout this project.

This project was supported by funding from the National Health and Medical Research Council, Australian Research Council, Australian Cancer Research Foundation, David Collins Leukaemia Foundation, Leukaemia Foundation of Australia and the Cancer Council Tasmania.

Abstract

Haematological malignancies (cancers of the haematopoietic and lymphoid tissues) are collectively one of the most frequently diagnosed cancers in Australia. Family history is one of the strongest risk factors for disease. Evidence for this derives from large population-based studies that have identified an increased risk of haematological malignancies in first degree relatives of cases, as well as studies of individual families where analyses have identified genes where family specific germline mutations predispose to these malignancies. Despite intensive research into the genetic predisposition to these cancers, the known genes account for only a small portion of the overall inherited component of haematological malignancies, leaving a significant gap in our understanding of the genetic basis of disease. Earlier studies used candidate gene approaches or sparse sets of genome wide markers to identify predisposition genes. Such approaches have a limited capacity for disease gene identification. Now, application of innovative technologies, such as next generation sequencing, to familial datasets with multiple cases of haematological malignancies presents an ideal opportunity to identify new predisposing germline mutations and other genetic factors contributing to disease development.

The aim of this study was to identify the genetic architecture of disease susceptibility in large families affected by multiple subtypes of haematological malignancies. This study takes advantage of a collection of extended Tasmanian haematological malignancy pedigrees comprising 48 families, as well as 84 additional Tasmanian haematological malignancy cases with no known family history of disease. This resource is particularly valuable due to the recognised stability and relative genetic homogeneity of the island population of Tasmania.

Next generation sequencing approaches were employed to identify novel, rare and shared predisposing mutations in affected family members. This was achieved through a combination of whole exome and whole genome sequencing in five prioritised families. Genome and exome alignment and variant calling were conducted using BWA and SAMtools. High-quality single nucleotide and small insertion / deletion variants identified were then annotated with information from public data

sources using ANNOVAR. Variants were filtered to focus in on rare variants (with population frequency estimates of 1% or less) using frequencies in Caucasian population data from the 1000 Genomes Project and the UK10K consortia dataset. A large number of rare shared genetic mutations were identified between related haematological malignancy cases in these families. A tiered prioritisation strategy was developed and employed to identify the top preferred candidates for further follow-up. This strategy incorporated variant-based prioritisation, using *in silico* predictions of variant effect, and gene-based prioritisation using known gene biology. For gene-based prioritisation a literature curated network analysis tool (Ingenuity Pathway Analysis) and an ontology-based tool (Phevor) as well as publically available tissue expression profiles of the mutated genes were used. Genes prioritised for further follow-up include examples such as *TNFSF9*, *TDP2*, *MMP8*, and *NOTCH1*. These genes have not been previously implicated in the familial risk for haematological malignancies, although some have previously established roles in malignancy. For example, *TNFSF9* is a gene with clear connections to both T-cell and B-cell biology and there is evidence from a mouse knockout model that disruption to this gene can contribute to malignancy development.

A subsequent aim of this study was to explore the role of telomere biology in familial haematological malignancies. Telomere biology has a well-characterised role in cancer development. Disruption of key telomere biology genes has been shown to lead to a spectrum of syndromes of which haematological malignancies are a feature such as dyskeratosis congentia and aplastic anaemia. To examine whether disrupted telomere biology was detectable in haematological malignancies, an analysis of telomere length was conducted using a PCR-based assay measuring across the familial resource, non-familial cases and population controls. Telomere length was analysed as a quantitative trait using variance components modelling, adjusting for age, sex and importantly kinship. The key finding from this analysis was that telomere length was highly heritable at 62.5% ($P=4.7\times10^{-5}$) indicating a strong genetic effect driving variation in telomere length and that both familial and non-familial haematological malignancy cases had shorter telomeres ($P=2.2\times10^{-4}$ and 2.2×10^{-5} respectively). These results indicate that telomere length contributes broadly to haematological malignancies. Genetic variation in some of the known telomere

biology genes was examined, however the underlying genetic contribution to the observed shortened telomere length remains to be determined.

This thesis describes the genetic analysis of a rare resource, providing evidence for several novel genes with possible roles in the development of haematological malignancies. As expected next generation sequencing of these families has further highlighted the multigenic contribution to risk in this complex disease.

Abbreviations and acronyms

1000GP	1000 Genomes Project
ABVD	Adriamycin, bleomycin, vinblastine, dacarbazine chemotherapy regimen
AIDS	Acquired immunodeficiency syndrome
AIHW	Australian Institute of Health and Welfare
ALL	Acute lymphocytic leukaemia
AML	Acute myeloid leukaemia
APML	Acute promyelocytic leukaemia
BDT	Big Dye Terminator
BL	Burkitt lymphoma
bp	Base pair
CADD	Combined Annotation Dependent-Depletion score
CGS	Candidate gene study
ChIP-Seq	Chromatic immuno-precipitation assay sequencing
cHL	Classical Hodgkin lymphoma
CLL	Chronic lymphocytic leukaemia
CML	Chronic myeloid leukaemia
COSMIC	Catalogue Of Somatic Mutations In Cancer
EBV	Epstein-Barr virus
ENCODE	Encyclopaedia Of DNA Elements
ET	Essential thrombocythaemia
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium
FDG-PET	Fluorodeoxyglucose positron emission topography
FL	Follicular lymphoma
FPD	Familial platelet disorder
GWAS	Genome-wide association study
GWLS	Genome-wide linkage study
HCL	Hairy cell leukaemia
HIV	Human immunodeficiency virus

HL	Hodgkin lymphoma
HM	Haematological malignancy
ISRT	Involved site radiation therapy
IPA®	Ingenuity® Pathway Analysis from QIAGEN
JMML	Juvenile myelomonocytic leukaemia
LP	Lymphoproliferative
LPL	Lymphoplasmacytic lymphoma
MAF	Minor allele frequency
MBCN	Mature B-cell neoplasm
MCL	Mantle cell lymphoma
MDS	Myelodysplastic syndrome
MDS-RARS	Myelodysplastic syndrome with refractory anaemia and ringed sideroblasts
MEN1	Multiple endocrine neoplasia type 1
MF	Myelofibrosis
miRNA	Micro RNA
MM	Multiple myeloma
MP	Myeloproliferative
MPD	Myeloproliferative disease
MPN	Myeloproliferative neoplasm
NCBI	National Center for Biotechnology Information
NCG 4.0	Network of Cancer Genes version 4.0
ncRNA	Non-coding RNA
NF1	Neurofibromatosis 1
NGS	Next generation sequencing
NHL	Non-Hodgkin lymphoma
NK	Natural killer
NLPHL	Nodule lymphocyte predominate HL
OR	Odds ratio
PCR	Polymerase chain reaction
pfam	Protein family
PV	Polycythaemia Vera
rpm	Revolutions per minute
SilVA	Silent Variant Analyzer

SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOLAR	Sequential Oligogenic Linkage Analysis Routines software
TFHMS	Tasmanian Familial Haematological Malignancies Study
TNF	Tumour necrosis factor
UCSC	University California Santa Cruz
UK	United Kingdom
USA	United States of America
UTR	Untranslated region
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organisation
WM	Waldenström macroglobulinemia

Table of Contents

Table of Figures

Table of Tables

Chapter 1 - Introduction

1.1 Haematological malignancies

Haematological malignancies (HMs) are cancers of the haematopoietic and lymphoid tissues and collectively are the third most diagnosed cancer in Australia and an important worldwide contributor to cancer related mortality[2,3]. An area of intense research is to identify why HMs occur, as this will lead to new mechanisms by which to prevent or treat disease. HMs, as for other types of cancers, are multifactorial diseases consisting of a complex interplay of genetic, environmental and lifestyle-based risk factors. One understanding of how cancers including HMs develop likens the process to evolution[4,5]. Normal cells become malignant through a multi-step process, whereby several genetic mutations are acquired over time that change the cells such that they acquire or evolve the 'hallmarks' of cancer cells[4,5]. These hallmarks and characteristics, as proposed by Hanahan and Weinberg[6,7], are summarised in Table 1.1. Together these hallmarks describe the complex barriers cells must overcome, by acquiring genetic mutations, to progress to malignancy. Cells evolve into a malignant state through development of several mutations in key genes that in sum result in acquisition of the hallmarks described in Table 1.1.

Table 1.1 Summary of the 'Hallmarks of Cancer' paradigm proposed by Hanahan and Weinberg, adapted from Hanahan *et al.*[7]

Established cancer hallmarks	**Mechanism**
Sustaining proliferative signalling	Recruit extra or increase sensitivity to growth factors
Evading growth suppressors	Inactivate tumour suppressor genes such as RB1 (retinoblastoma 1) and TP53 (tumour protein p53)
Enabling replicative immortality	Upregulated expression of telomerase, or alternate lengthening of telomeres, which prevents the triggering of cellular senescence or cellular crisis (and the subsequent cell death)
Activating invasion and metastasis	Loss of cell-to-cell adhesions, inactivation of E-cadherin, upregulation of N-cadherin, activation of EMT (epithelial-mesenchymal transition)
Inducing angiogenesis	Upregulated proangiogenic signals such as VEGFs (vascular endothelial growth factors) and FGFs (fibroblast growth factors) to form new vascularisation into the tumour site
Resisting cell death	Limit or prevent apoptosis by loss of TP53 function, increasing expression of anti-apoptotic factors (e.g. Bcl-2), decreasing expression of pro-apoptotic factors (Bax, Bim, Puma)
Emerging hallmarks	**Mechanism**
Avoiding immune destruction	Suppression of immune system, production of cancer cells that do not trigger an immune response
Deregulating cellular energetics	Limit energy supply of cancer cells to aerobic glycolysis, enhancing proliferative capacity
Enabling characteristics	**Mechanism**
Tumour-promoting inflammation	Inflammation provides bioactive molecules such as growth factors and proangiogenic factors to the developing tumour
Genome instability and mutation	Acquisition of successive genetic alterations (acquired mutations) for further hallmarks to develop, deregulation of DNA repair mechanisms

1.2 Deregulation of haematopoiesis leads to malignancy

HMs arise as a result of deregulated haematopoiesis. Haematopoiesis is the biological process by which the mature cells of the haematopoietic and lymphoid tissues are formed. In healthy individuals haematopoiesis occurs in the bone marrow, where niches of long-term haematopoietic stem cells form short-term haematopoietic stem cells, from which the common myeloid and common lymphoid progenitors derive[8,9]. These progenitors can differentiate via a number of pathways to form the cellular components of blood as shown in a simplified model in Figure 1.1. Briefly, the common myeloid progenitors form two branches, the megakaryocyte / erythroid progenitors, which in turn form mature red blood cells and platelets, and the granulocyte / macrophage progenitors, which form mast cells, neutrophils, monocytes and macrophages, dendritic cells, eosinophils and basophils[8,9]. The common lymphoid progenitors form B-cells and plasma cells, T helper cells, cytotoxic T-cells and natural killer (NK) cells[8,9].

Haematopoiesis is not solely a linear process and it has been suggested that the simplified haematopoiesis hierarchical model as depicted in Figure 1.1 should be interpreted as an illustration of the differentiation possibilities[8,10]. There is also *in vitro* evidence showing that differentiation lineages can be reprogrammed (by forced expression of targeted transcription factors) from one lineage to another. But whether this can occur *in vivo* has yet to be determined[11-13]. Figure 1.2 adapted from Ceredig *et al.*[10] illustrates an alternate model of haematopoiesis supporting the concept of a non-linear process whereby multiple different types of progenitors can differentiate to form the same terminally differentiated mature cells.

Further underlining the idea that haematopoiesis is a system of differentiation possibilities is the observation of age related differences in haematopoietic stem cells. Studies have shown that with age, haematopoietic stem cells change, and become unstable and biased towards myeloid lineages over lymphoid lineages, and that this is in parallel with changes in gene expression with increased expression of malignancy-related genes and decreased expression of DNA damage repair genes[14-17]. Indeed the specific transcription factor requirements of the different lineages in haematopoiesis are intimately linked with the potential for disruption to occur and HMs to form, with

the type of HM developing depending upon where and in which lineage disruption occurs.

Key regulatory genes that are essential for haematopoiesis have been shown to be involved in the development of HMs[8]. An example is *RUNX1* which is an essential transcription factor for haematopoiesis and haematopoietic stem cell formation[8,18], and is regularly disrupted often by chromosomal translocations in a range of leukaemias[19-21]. Another example is the transcription factor *GATA1* which is required in the myeloid lineage in haematopoiesis and is also causally involved in the HMs associated with Down syndrome[22,23]. These and many other genes essential for haematopoiesis can contribute to HM development when aberrantly regulated. What subtype of HM develops is dependent upon where in haematopoiesis and specifically what lineage is affected.

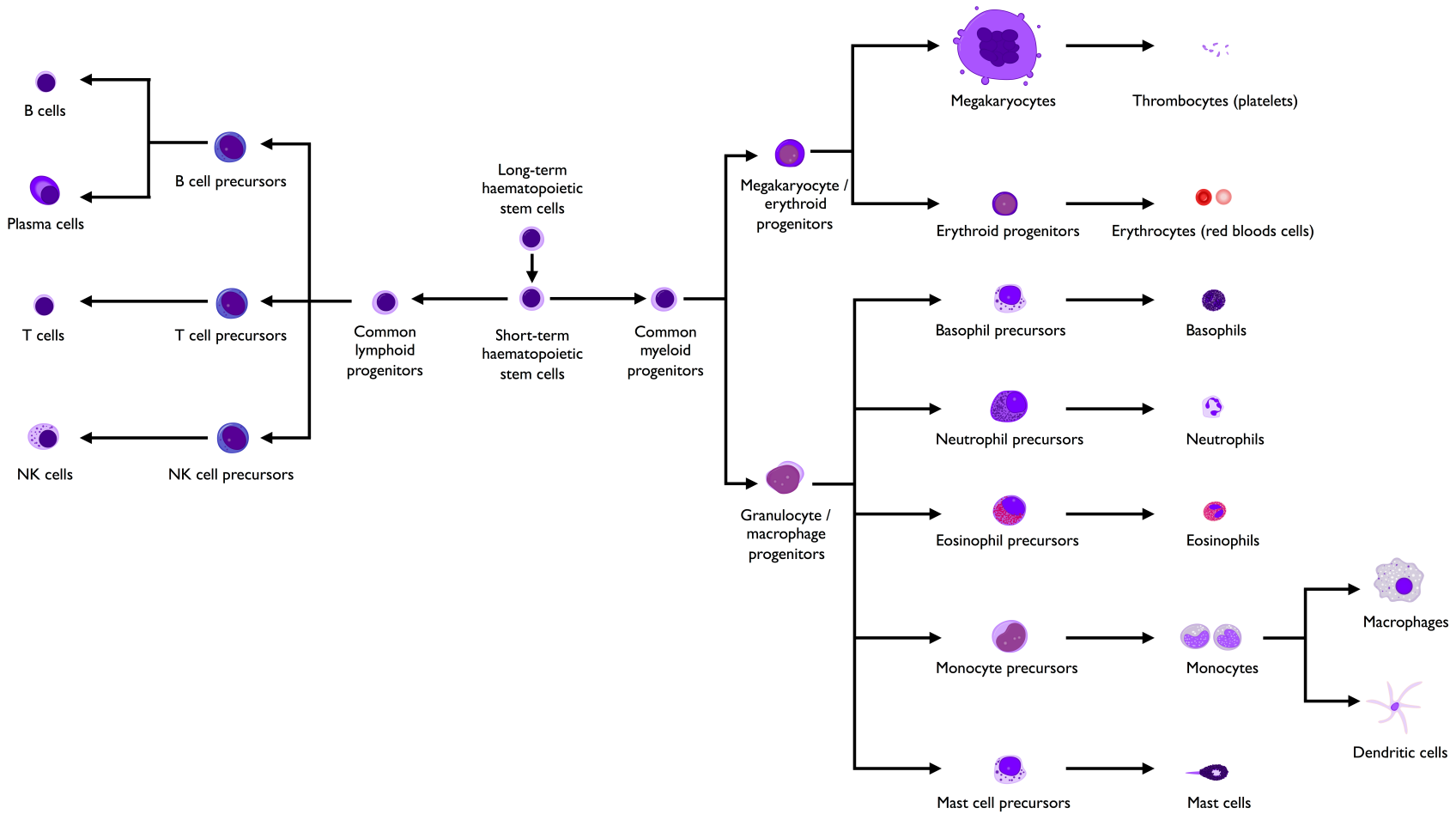


Figure 1.1 A simplified hierarchical model of haematopoiesis.
Self-renewing haematopoietic stem cells differentiate into mature cells of either the lymphoid or myeloid lineage. Figure constructed using images adapted under the creative commons license from http://commons.wikimedia.org/wiki/File:Haematopoiesis_(human)_diagram.png by A. Rad and by adaptation of information in Orkin *et al.*[8], Ceredig *et al.*[10] and Hatton *et al.*[9].

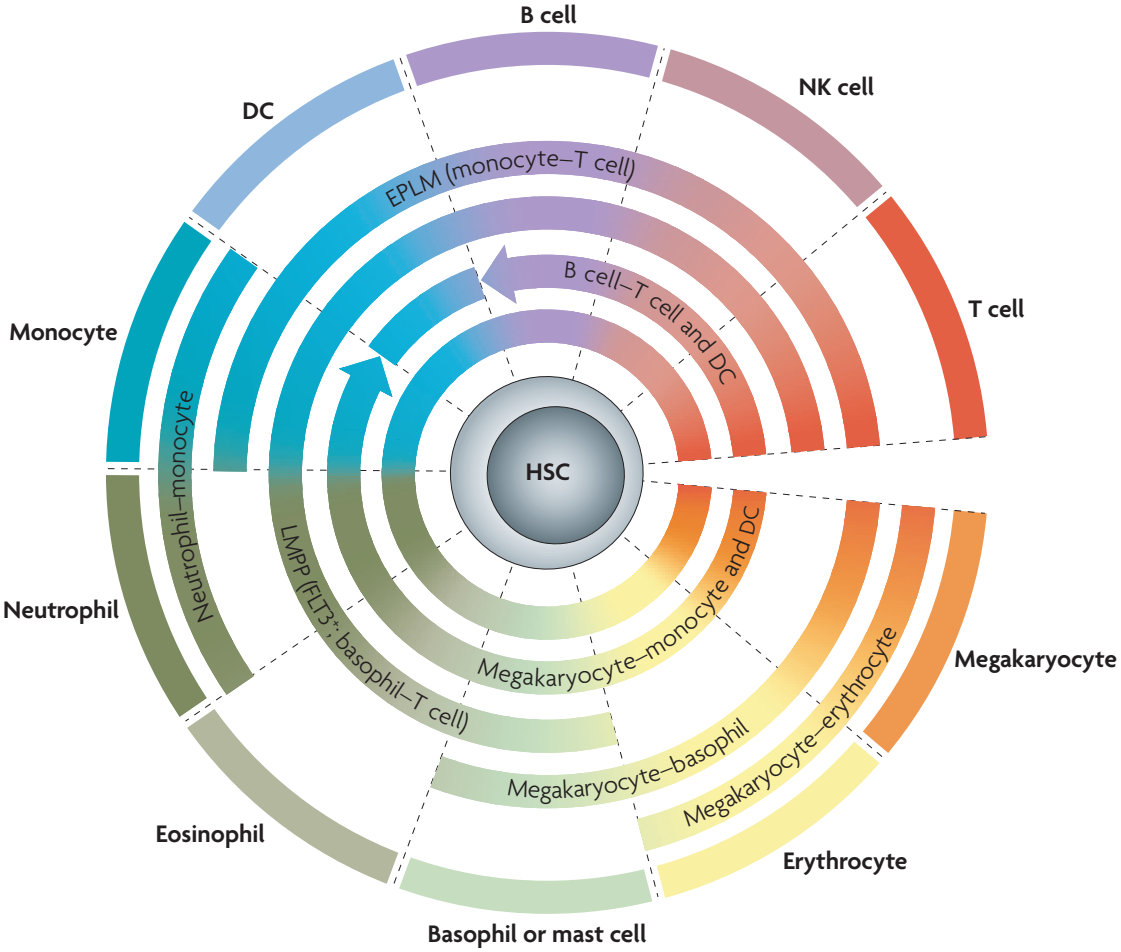


Figure 1.2 An alternate model of haematopoiesis from Ceredig *et al.*[10]
This model, as proposed in Ceredig *et al.*[10], shows haematopoiesis as a continuum of differentiation with progenitor cells having flexibility to move between related differentiation lineages, in contrast to the hierarchical commitment in Figure 1.1. As described[10] this model makes no assumptions as to the hierarchical structure of differentiation instead showing arcs that indicate known oligopotent progenitor cells, e.g. the Megakaryocyte-erythrocyte arc shows the potential of these progenitors to form both megakaryocytes and erythrocytes. DC, dendritic cell; EPLM, early progenitor with lymphoid and myeloid potential; FLT3, FMS-related tyrosine kinase; LMPP, lymphoid-primed multipotent progenitor; NK cell, natural killer cell.

1.3 Population and clinical aspects of haematological malignancies

Broadly HMs can be stratified into three major groups: leukaemias, lymphomas, and myelomas. Within these the World Health Organization (WHO) has classified over 100 different specific subtypes of HMs on the basis of morphology, immunophenotype, genetic features and clinical features[24]. Subtypes vary widely in their diagnosis, treatment and clinical outcomes[24]. What HM subtypes do share in common is that they arise from genetic abnormalities that affect haematopoiesis. Myeloid malignancies such as acute myeloid leukaemia (AML), chronic myeloid leukaemia (CML) and myeloproliferative neoplasms (MPNs) arise from deregulation of the myeloid lineage[24]. In the same way, lymphoid malignancies such as acute lymphocytic leukaemia (ALL), chronic lymphocytic leukaemia (CLL), multiple myeloma (MM) and Hodgkin lymphoma (HL) arise from deregulation of the lymphoid lineage[24].

1.3.1 Population impact of haematological malignancies

The Australian Institute of Health and Welfare (AIHW) has shown consistently from 2003 to 2010 that the number of HMs diagnosed yearly in Australia exceeds 10,000 people[25]. According to 2010 AIHW data, HMs, with 11,559 cases nationally, were the third most common cancer after prostate and breast cancers[25]. HMs are reported across 16 different classes in AIHW statistics. Table 1.2 shows the AIHW 2010 cancer incidence data for the top six most frequently diagnosed cancers, with the 16 classes of HMs combined to show the total incidence and then with the specific class breakdown. Figure 1.3 shows the incidence relationships of these cancers in Australia using AIHW data from 1982 to 2010.

Table 1.2 AIHW 2010 data for the top 5 cancers diagnosed nationally, with haematological malignancies consisting of 16 separate AIHW classification classes[25].

AIHW Cancer Classification	Male	Female	Total
C61 Prostate	19,821	0	19,821
C50 Breast	127	14,181	14,308
Haematological malignancies	**6,662**	**4,897**	**11,559**
C81 Hodgkin lymphomas	324	248	572
C82 Follicular non-Hodgkin lymphoma	542	446	988
C83 Diffuse non-Hodgkin lymphomas	1,229	848	2,077
C84 Peripheral and cutaneous T-cell lymphomas	168	95	263
C85 Other and unspecified non-Hodgkin lymphomas	605	529	1,134
C88 Immunoproliferative cancers	53	38	91
C90 Multiple myeloma and other plasma cell cancers	826	641	1,467
C91 Lymphoid leukaemias	952	600	1,551
C92 Myeloid leukaemias	792	609	1,402
C93 Monocytic leukaemias	36	25	61
C94 Other leukaemias of specified cell type	15	16	30
C95 Leukaemias of unspecified cell type	35	31	65
C96 Other and unspecified cancers of lymphoid, haematopoietic and related tissue	17	20	37
D45 Polycythaemia vera	101	90	191
D46 Myelodysplastic syndromes	764	476	1,241
D47 Other cancers of lymphoid, haematopoietic and related tissue	204	185	389
C43 Skin - Melanoma	6,700	4,705	11,405
C34 Bronchus and lung	6,241	4,042	10,283

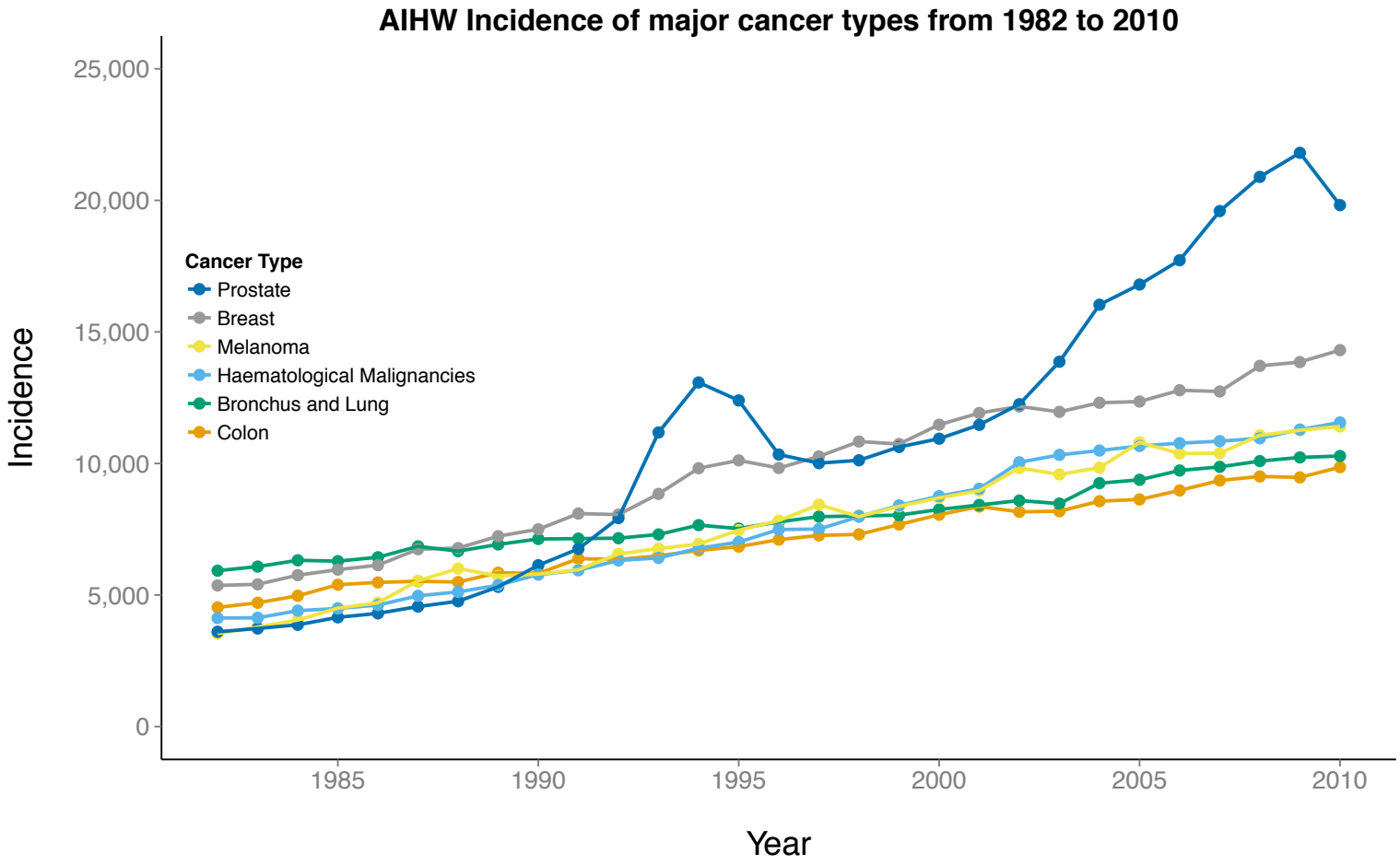


Figure 1.3 Incidence of major cancer types in Australia from 1982 to 2010 using data from the Australian Institute of Health and Welfare[25].
This figure shows that consistently, HMs are one of the most frequently diagnosed cancers in Australia.

GLOBOCAN 2012[3] is the most recent comprehensive worldwide study of cancer in the adult population (ages 15 and over). GLOBOCAN reports that the estimated 5-year population prevalence of HMs worldwide (categorised as leukaemia, multiple myeloma, Hodgkin lymphoma and non-Hodgkin lymphoma) is 33.7 per 100,000 people[26]. GLOBOCAN 2012 estimates the 5-year population prevalence of HMs in Australia is 134.9 per 100,000 people. This is similar to other 'Western' countries such as the United States of America and the United Kingdom with 5-year population prevalence estimated at 134.3 and 120.1 per 100,000 people respectively. These country specific figures are much greater than GLOBOCAN's estimated worldwide prevalence of 33.7 per 100,000, as there are countries with much lower prevalence estimates.

The difference in prevalence between the worldwide figure and the Australian/USA/UK figures is due to the much lower 5-year prevalence of HMs in regions such as Africa (15.8), Asia (17.2) and South America (28.3). This is illustrated in Figure 1.4, four world maps generated from GLOBOCAN 2012 with the prevalence of HMs across the world described as heat maps. As can be seen in Figure 1.4 there is a trend that Caucasian countries, predominantly the UK, USA and Australia, have a much higher 5-year prevalence of HMs in comparison to African and Asian countries. As HMs are multifactorial complex diseases this immediately questions whether region specific environmental risk factors contribute to the higher prevalence of HMs within particular countries or whether genetic ancestry drives this prevalence disparity. GLOBOCAN worldwide data may also be influenced by incomplete reporting rates and low quality data from non-Western country regions.

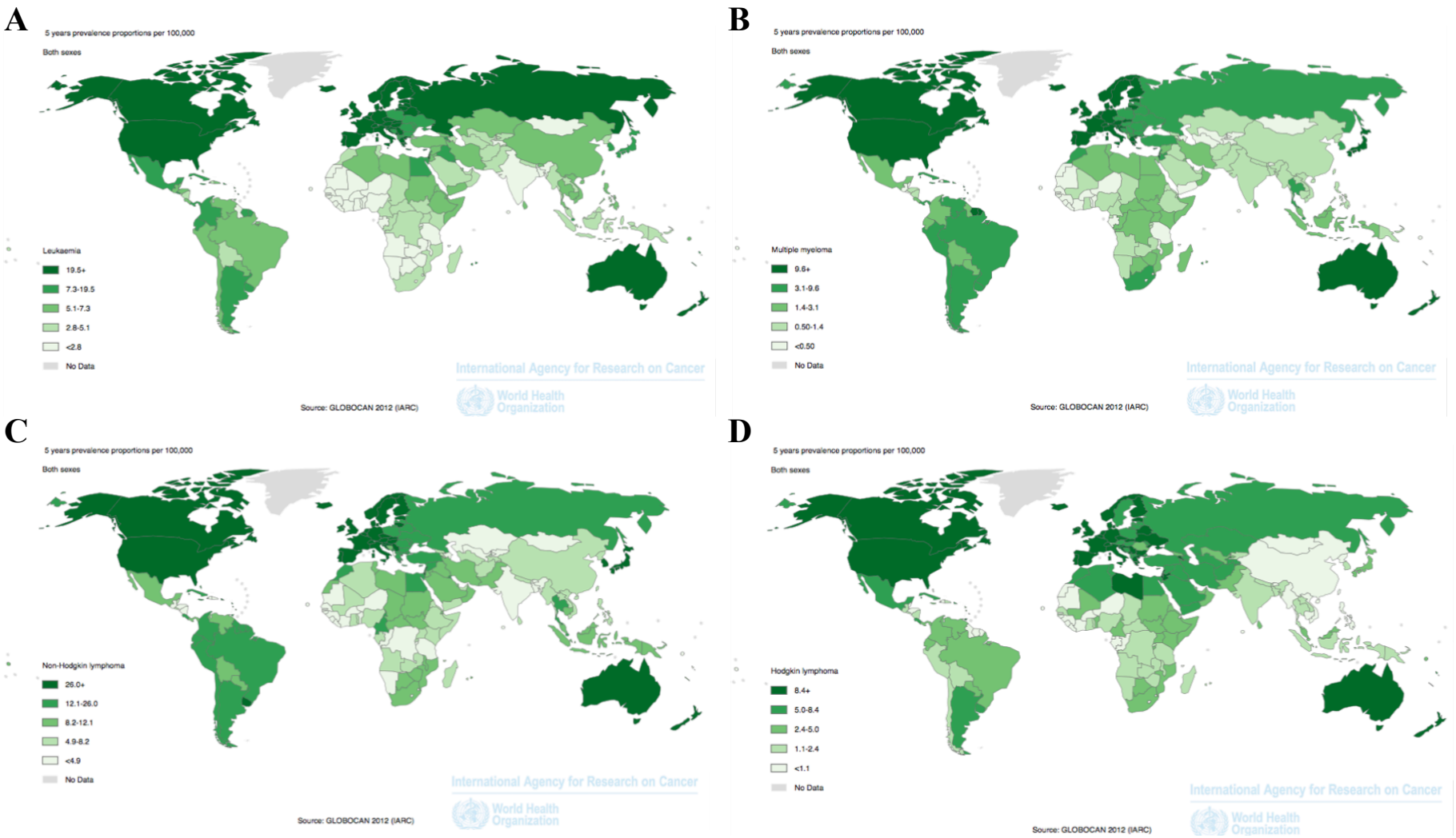


Figure 1.4 GLOBOCAN 2012 worldwide prevalence of HMs per 100,000 people.
A) Leukaemia, B) multiple myeloma, C) non-Hodgkin lymphoma, D) Hodgkin lymphoma. Obtained from http://globocan.iarc.fr[3].

1.3.2 Clinical aspects of haematological malignancies

Haematological malignancies recapitulate Hanahan and Weinberg's 'hallmarks of cancer' listed in Table 1.1[7]. Through acquisition of genetic mutations, the normal process of haematopoiesis becomes a source of malignancy as cells obtain replicative immortality with sustained proliferation and avoidance of the immune system and apoptosis. HMs may present as primarily a circulating disease within the blood, such as observed in CLL, or in other subtypes solid tumours are the primary feature of disease, for example HL[24]. Some HMs are asymptomatic in early stages of disease, and may be first detected as a result of standard general practice medical examinations, presenting as abnormal blood cell counts[24].

1.3.3 Challenges of therapies in haematological malignancies

One of the major goals of HM therapies is to reduce toxicity and increase survival. For younger patients, for example childhood ALL, therapeutic strategies must weigh disease eradication against the potential for long-term side effects of therapy, such as therapy related malignancies[27,28]. In older patients, many suffer from additional comorbidities that reduce fitness for intensive therapy, which also influences therapy choice. For many HMs prognostic genetic factors that assist in stratifying patients into treatment groups have been identified. For example in CLL a recurring genetic factor is deletions of the p arm of chromosome 17 (encompassing the tumour suppressor gene *TP53*). Patients with del(17p) do not respond as well as other CLL subgroups to combination chemotherapy with fludarabine and rituximab, so other treatment options such as allogeneic stem cell transplant are used[29,30].

HMs have a high burden across the population, so identifying ways to improve treatment results and minimise later complications arising from therapy is a priority research area. One aspect of this is identifying factors that can be used to predict disease risk, progression or treatment response.

1.4 Risk factors for haematological malignancies

As a multifactorial disease, a number of risk factors have been associated with the development of HMs. A strong risk factor for HMs is a family history of disease and

accordingly the primary topic of this thesis is understanding the genetic basis of disease. However it is also important that genetic risk factors are considered in light of known non-familial and environmental risk factors.

1.4.1 Non-familial risk factors

1.4.1.1 Age

Increased age is a well-known risk factor for cancer and increasing age is associated with cancer risk[31]. This is reflected in HMs where data from the AIHW show that the mean age of disease incidence is above 60 years across different HM subtypes[2]. The exception is HL where the mean age of disease incidence is 40[2]. Even so, some subtypes of HMs are considered clinically to be childhood diseases. AML, for example, is typically a childhood cancer, but can also be seen in adults[24]. Whereas subtypes such as CLL and the MDS are rare in children and are seen more commonly in adults[24].

The risk of developing HMs increases with age with risk magnitudes differing between subtypes. For example in Australia the lifetime risk of developing CLL to age 75 is one in 304 people with this risk almost doubling to one in 161 if the age range is extended to age 85[2]. For HL the mean age of incidence is around 20 years younger than CLL however the overall risk is lower being one in 549 to age 75 and one in 463 to age 85[2]. Therefore age as a risk factor is important in HMs but the risk effect varies considerably from subtype to subtype.

Further, the increased risk of cancer with age has been attributed to the accumulation of acquired genetic mutations over time. However the logic of this axiom has been challenged recently by DeGregori who makes a compelling argument that as the rate of acquired genetic mutations peaks in ontogeny, why then does the incidence of cancer peak much later in life and not concurrently with mutation rate?[32] Thus genetics may be involved in the age-related increase in the incidence of HMs but whether this is solely due to the acquisition of genetic mutations over time remains to be determined.

1.4.1.2 Sex

Sex is another well-characterised risk factor for cancer, where broadly, the incidence of non-sex specific cancers is greater in men than women[2,31,33-35]. In Australians the incidence and mortality of HMs is greater in men than women across the age and subtype spectrum[2]. The exception is again HL where incidence changes with age, and is higher for women than men from around age 20 to 30 and again at around ages 50 and 75[2]. This age and sex interaction on risk occurs in other HM subtypes[36]. Speculation on the biological causes underlying the sex difference in cancers such as HMs has suggested factors such as differences due to hormones, genetics, epigenetics and immune surveillance, however the causes remain uncharacterised[37,38].

1.4.1.3 Environmental and lifestyle factors

A range of environmental and lifestyle factors also contribute to cancer development[5]. Factors such as smoking, occupational exposure to carcinogens, chronic infections and ionizing radiation, can cause additional damage to haematopoietic cells that additively contributes to a transformation to malignancy[5,39]. A US study of population birth cohorts showed that the incidence of leukaemia varied significantly across birth years and per generation[40]. In this study the incidence of AML and ALL was seen to increase across the temporal cohort spectrum, while CML and CLL declined. This could indicate a change in environmental exposures and lifestyle factors resulting in increased incidence of AML and ALL and decreased incidence of CML and CLL. These incidence differences between birth cohorts support the hypothesis that adult leukaemia is modulated by environmental and lifestyle exposures and that these exposures change over time across cohorts.

In the World Cancer Report 2008, the International Agency for Research on Cancer listed a range of industries where occupational exposure to particular carcinogens was attributed to increasing the incidence of leukaemias and lymphomas in workers of those industries (as summarised in Table 1.3, data drawn from[39]). One occupational carcinogen example is the solvent benzene, which is also present in cigarette smoke. Studies have shown an association between benzene exposure from smoking and the incidence of myeloid leukaemias[41] and exposure due to occupation and CLL and MM incidence[42,43]. Even though there is a clear involvement between benzene and HM

incidence the association has differing levels of evidence between different subtypes and is compounded by other co-occurring occupational exposures[44]. Identifying what is and what is not an occupational risk factor for HMs is challenging due to compounding factors, difficulties with sample size and quantifying exposures but is still crucial for disease management and prevention.

Table 1.3 Occupational exposures to environmental risk factors believed to contribute to the development of HMs. Data from the World Cancer Report 2008[39].

Carcinogen	Industry / Industrial use	HM type
Benzene	Solvents, fuel	Leukaemia
1,3-Butadiene	Plastics and rubber industries	Leukaemia
Non-arsenical insecticides	Agriculture	Leukaemia
Polychlorinated biphenyls	Electrical components	Lymphoma
Tetrachloroethylene	Solvent	Lymphoma
Trichloroethylene	Solvent, dry cleaning	Lymphoma
	Boot and shoe manufacture and repair	Leukaemia
	Petroleum refining	Leukaemia

Ionizing radiation is a well-established carcinogen that contributes to the risk of HM development. Exposure can occur occupationally, medically or due to close proximity to events and locations such as Chernobyl and the atomic bombings of Japan. It has been proposed that ionizing radiation can act directly on stem cells in the body, including haematopoietic stem cells, inducing changes that increase the risk of malignancy development[45]. Nakamura suggests that ionizing radiation could be acting on pre-malignant cells randomly adding additional genetic mutations which in turn trigger cellular progression to malignancy[46]. This is particularly in line with early research that showed that pre-natal exposure to X-ray irradiation during obstetric examination raised childhood cancer rates, including leukaemia[47,48].

Retrospective studies of Japan's atomic bomb survivor cohort have also drawn the link between ionizing radiation and the increased occurrence of HMs, particularly AML[49,50], but also MDS[51] with risk of developing some subtypes persisting up to five decades after the initial exposure. However, an increased risk of CLL, HL and MM were not detected in this cohort[50]. Other studies of CLL and other leukaemias outside of Japan have reported conflicting results as to whether there is an association between disease risk and ionizing radiation exposure. For example, studies of cohorts

in Ukraine[52] and Russia, Belarus and the Baltic countries[53] show that exposure to ionizing radiation is associated with an increased risk of HMs including CLL. Whereas other cohort studies have shown that there is an increased risk for leukaemia development in those exposed to radiation, except for CLL[54,55].

Virus infection is another environmental / lifestyle factor that has been shown to contribute to the development of HMs. Adult T-cell leukaemia is strongly related to infection by human T-cell leukaemia virus type-1, particularly in Japan where there is an endemic cluster of this HM subtype[56-58]. Patients infected with human immunodeficiency virus (HIV) have an increased risk of lymphoma development due to progressive immune system suppression, particularly in the later stages of acquired immunodeficiency syndrome (AIDS)[59-61]. Epstein-Barr virus (EBV), as well as being an additional factor in the development of HIV/AIDS related lymphoma, has been shown to independently related to the development of HL, Burkitt lymphoma (BL) and non-Hodgkin lymphoma (NHL), as reviewed in Kutok *et al.*[62].

1.4.1.4 Ancestry / ethnicity

As discussed in section 1.3, there are significant differences globally in the prevalence of HMs when looking at data from GLOBOCAN 2012[3]. Asian, African, Russian and Baltic countries have a lower prevalence of HMs than 'Western countries' including Northern and Western Europe, Northern America (excluding Mexico) and Australia as illustrated globally in Figure 1.4. One possible explanation is that the differences between high and low prevalence countries are due to environmental and lifestyle factors. There is evidence to support this in specific subtypes of HMs such as BL resulting from EBV infection in Africa[63], however research also points to ethnicity and genetic ancestry as the main driver of these differences. It is also possible that limited data collection and diagnoses from developing Asian and African countries drives these geographical differences. Studies assessing migrant populations in countries with high incidences of HMs can overcome the data collection problem, allowing a focus on differences between ethnicities. For example, studies of migrant Asians in America have shown that the migrant groups retain the low HM prevalence of their ethnic background[64]. This lower prevalence is maintained across subsequent generations born in the high incidence country[64]. Similar findings came from a data-

linkage study in the UK where HM prevalence in British Caucasians was compared to multiple UK migrant ethnic groups[65]. These studies support a role for genetics in the ethnicity differences in HM prevalence and are evidence against ethnicity specific environmental and lifestyle factors being the primary difference.

It is also possible that the same HM subtype is fundamentally different in different ethnic groups. However a recent study compared the molecular profiles of Asian and Caucasian CLL cases[66], finding the profiles to be much the same. This shows that at least for CLL there is not a fundamental difference in the cancer genomes between Asian and Caucasian disease, yet it does not explain why CLL is more prevalent in Caucasians. This study did not compare the germline genome of the cases. It is likely that the differences in HM prevalence between ethnicities is due to the underlying genetic differences between groups and points to an inherited basis of disease. Therefore ethnicity and ancestry have an important role in the development of HMs but our understanding remains limited with further work required to tease out specific ethnicity risk factors for HMs.

Moving away from the global population level consideration of HMs, what these findings from ethnicity studies do suggest is that there are inherited risk factors that have an important role in HM development. Different types of studies can address components of these ethnicity and inheritance-based questions.

1.4.2 Familial risk in haematological malignancies

1.4.2.1 Population-based studies of familial HMs

One of the earliest reports providing population-based evidence of a familial risk for HMs is from 1947 when Videbæk published his pioneering work comparing Scandinavian families of 209 leukaemia probands to the families of 200 unaffected individuals[67]. This publication was one of the first to signal a change in the field away from an environmental or contagious cause of HMs towards a genetic basis for disease. Videbæk's work revealed that in a series of 209 patients with leukaemia, 17 had evidence of familial leukaemia whereas in the relatives of the 200 unaffected control individuals there was only one case of leukaemia. At the time this was the most extensive and comprehensive study of familial HMs that went beyond single

case reports. It was also one of the first suggestions that multiple subtypes of HMs, as well as other cancers, could occur in the same family and may have a genetic connection.

In more recent years, large studies have repeatedly shown that HMs aggregate in families and that a family history of disease is a significant risk factor for HMs. A familial risk for HMs is supported by the results of large epidemiological studies that link data in the Swedish Cancer Registry[68] and the Swedish Multigenerational Registry[69]. Together these registries provide information on cancer patients and their relatives in Sweden back to 1958. As summarised in Table 1.4 and Figure 1.5 most studies have shown that first-degree relatives have a statistically significant increased risk of developing the same HM subtype as the related case. Additionally, most studies have shown that first-degree relatives of HM cases have a statistically significant increased risk of developing other subtypes of HMs. For example one study has shown that first degree relatives of CLL cases have an 8.5 fold increased risk of developing CLL and increased risks for other subtypes of NHLs including B-cell NHL, hairy cell leukaemia (HCL), and lymphoplasmacytic lymphoma / Waldenström macroglobulinemia (LPL/WM)[70]. By identifying an increased risk in first-degree relatives of HM cases these epidemiological studies provide compelling evidence of an inherited genetic risk for HMs. Similar studies across a number of centres, using the Danish Cancer Registry[71], Swedish Cancer Registry[72-74] and studies from the United Kingdom[75], Utah[76] and a multi-centre consortium study[77] support these findings.

Table 1.4 Summary of the findings from population based studies of familial HMs.

Proband Disease	LPL/WM[78]	MPN[79]	CLL[70]	DLBCL[80]	FL[80]	HL[80]	MM[81]	MGUS[82]	AML[83]	MDS[83]
Study Details										
Cases	2144	11039	9717	2517	2668	6963	13896	4458	6962	1388
# 1st degree case relatives	6177	24577	26947	8974	10188	24053	37838	14621	20579	3994
Controls	8279	43550	38159	9932	10468	28371	54365	17505	27872	5312
# 1st degree control relatives	24609	99542	107223	35310	40384	108180	151068	58387	90406	15818
Period	1958-2005	1958-2005	1958-2004	1958-2004	1958-2004	1958-2004	1958 - 2005	1967-2005	1958-2004	1993-2004
Relative Risk in First Degree Relatives RR (95% C.I.) with * signifying statistical significance in that study, NR = not reported										
MM	1.6 (0.8–3.2)		1.2 (0.85–1.8)				**2.1 (1.6–2.9)***	**2.9 (1.9–4.3)***	1.0 (0.6–1.5)	0.7 (0.2–2.0)
MGUS	**5.0 (1.3–18.9)***		1.4 (0.88–1.8)				**2.1 (1.5–3.1)***	**2.8 (1.4–5.6)***		
ALL							**2.1 (1.0–4.2)***		1.2 (0.5–2.7)	2.0 (0.2–21.3)
NHL	**3.0 (2.0–4.4)***		**1.9 (1.5–2.3)***				1.1 (0.8–1.7)	1.1 (0.8–1.5)	1.0 (0.8–1.3)	0.7 (0.3–1.4)
B-cell NHL			**1.8 (1.3–2.5)***	1.63 (NR)	1.81 (NR)	**1.54* (NR)**				
T-cell NHL			1.3 (0.36-4.9)							
Indolent B-cell NHL			**2.2 (1.5–3.2)***	1.66 (NR)	**2.11 (1.1–3.9)***	1.26 (NR)				
Aggressive B-cell NHL			1.0 (0.41–2.5)							
DLBCL				**9.85 (3.1–31)***		**2.45* (NR)**				
Mantle Cell Lymphoma			1.1 (0.24–5.5)							
Hairy Cell Leukaemia			**3.3 (1.0-10.9)***							
FL			1.6 (0.87–2.8)		**4.0 (1.6–9.5)***	1.35 (NR)				
CLL	**3.4 (1.7–6.6)***	**1.6 (1.1–2.5)***	**8.5 (6.1–11.7)***		**1.8 (1.0–3.3)***		1.1 (0.8–1.7)	**2.0 (1.2–2.3)***	**1.6 (1.0–2.5)***	0.5 (0.1–2.1)
LPL/WM	**20.0 (4.1–98.4)***		**4.0 (2.0–8.2)***				1.4 (0.7–2.8)	**4.0 (1.5–11)***		
HL	0.8 (0.3–2.2)		1.5 (0.96–2.3)	**2.04 (1.05–4.0)***	1.42 (NR)	**3.93* (NR)**	0.9 (0.6–1.4)	<45yo 1.3 (0.6–2.9), >45yo 0.2 (0.0–1.7)	1.3 (0.8–2.2)	0.8 (0.2–3.6)
Any LP (NHL, HL, CLL, MM)			**2.6 (2.2–3.0)***						1.1 (0.9–1.4)	0.7 (0.4–1.1)
AML/MDS							0.8 (0.5–1.2)		1.0 (0.6–1.9)	1.5 (0.5–4.8)
AML		1.3 (0.9–1.9)							0.9 (0.5–1.9)	1.1 (0.30–3.8)
MDS		1.6 (0.6–4.4)							1.8 (0.6–5.7)	4.0 (0.4–43.1)
CML		1.9 (0.9–3.8)					0.5 (0.2–1.2)		1.3 (0.6–2.9)	2.0 (0.2–21.6)
Any myeloid malignancy									1.1 (0.7–1.8)	1.6 (0.6–4.5)
MPD							1.1 (0.7–1.6)			
PV		**5.7 (3.5–9.1)***							**2.3 (1.2–4.5)***	0.8 (0.1–6.7)
ET		**7.4 (3.7–14.8)***							1.0 (0.4–2.6)	
MF		0.9 (0.2–4.2)							1.0 (0.4–3.1)	1.3 (0.1–12.6)
MPN unclassifed		**7.5 (2.7–20.8)***							1.1 (0.3–3.9)	
Any MPN (PV, ET, MF, MPD)		**5.6 (3.8–8.2)***							1.5 (0.9–2.3)	1.6 (0.50–5.0)
Any myeloid or MPN									1.3 (0.9–1.8)	1.6 (0.7–3.6)
Any HM									**1.2 (1.0–1.4)***	0.9 (0.6–1.3)

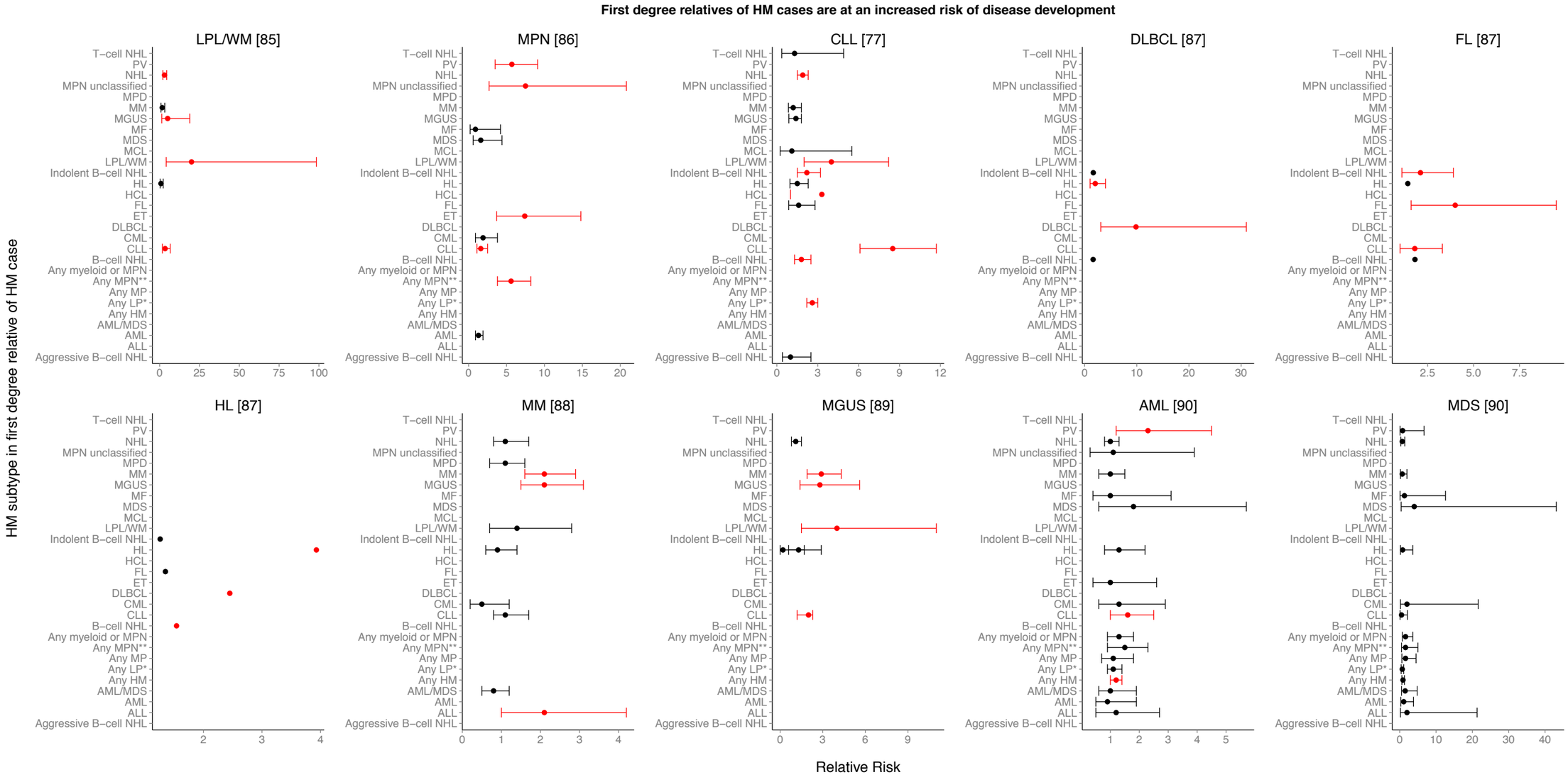


Figure 1.5 Summary of the findings from population based studies of familial HMs.
Relative risks for each study reported in Table 1.4 are graphed for comparison. Due to variation between studies in HM subtypes investigated, not all subtypes have a relative risk score. One study reported relative risk values without confidence intervals (DLBCL, FL and HL[80]). The relatives of MGUS patients[82] who were diagnosed with HL have two relative risk points corresponding to risk calculations in two age groups < 45 years old: 1.3 (0.6-2.9) and > 49 years old 0.2 (0.0-1.7). Statistically significant relative risks ($P < 0.05$) are indicated in red in the plots for each study. Each study is plotted on a different x-axis scale as indicated.

1.4.2.2 Evidence of familial risk from single family studies of familial HMs

A key omission in the population-based studies of familial HM risk is that they only examine whether first-degree relatives of HM cases are at an increased risk of disease, without extending to more distant relationships. Reports of individual families where HMs occur across multiple generations and multiple individuals indicate that the increased risk extends beyond that of first-degree relatives. Such families include those for which inherited syndromes are responsible for the HM predisposition, including Bloom syndrome[84,85], and dyskeratosis congenita[86-88], as well as HMs occurring in non-syndrome HM families[89-92]. Further evidence for the familial risk of HMs includes twin studies that have shown in monozygotic twins that if one twin develops a HM the second is also at an increased risk of developing a HM[93,94].

Typically more recent studies of familial HMs have focused on families with one subtype of HM with incidental recruitment of other subtypes whereas the early reports of familial HMs by Videæk and Gunz, show the occurrence of multiple subtypes in the same family[67,95]. Recent studies have recruited families without a subtype bias and report families with multiple subtypes of HMs[92,96,97]. Such reports, together with the large population studies of familial HMs support the hypothesis that there is an inherited risk of HM development across a range of HM subtypes and that the familial risk for HMs extends beyond first-degree relatives. Biologically this multiple subtype predisposition is intuitive in the context of Figures 1.1 and 1.2 where early haematopoietic progenitor cells can give rise to multiple mature cell lineages. Thus an inherited susceptibility may affect multiple related haematopoietic lineages giving rise to different HM subtypes in the same family.

1.5 Identifying the mechanism of familial risk in HMs

The underlying mechanisms behind the occurrence of HMs in families remain largely uncharacterised. Segel and Lichtman[98] proposed that the occurrence of HMs in families, and the increased familial risk for HMs, results from either chance, or the inheritance of predisposing genetic variants that increase the likelihood of developing cancer. Given the breadth of familial studies in this area, at both the population and individual family levels, it is unlikely that the observed aggregation of HMs in families is due to chance. Familial studies strongly support the idea that germline inheritance, be it through single mutations in cancer susceptibility genes or multiple predisposing mutations, has an important role in the development of familial HMs. Therefore genetic studies of families with clusters of HM subtypes will assist in the elucidation of mechanisms underlying the familial aggregation. Such an approach has proven successful previously for studies of other familial cancers such as breast cancer and Lynch syndrome[99-106].

For many cancers there exists a genetic predisposition to disease that precedes, and sometimes underpins, the acquired somatic mutations cancers develop during malignancy progression. While in recent years the focus of cancer genetics has been primarily on identifying somatic mutations, as evidenced by large scale projects such as The Cancer Genome Atlas (TCGA)[107] and the International Cancer Genome Consortium as databased in the Catalogue Of Somatic Mutations In Cancer (COSMIC)[108-110,377], an area of equal importance has been the identification of germline mutations in genes that predispose to cancer development. Such cancer predisposition genes, as defined in[111], are genes where rare mutations confer an increased risk of cancer in mutation carriers. Functionally we categorise these genes into two categories: tumour suppressor genes and oncogenes.

Tumour suppressor genes have an active role in preventing malignancy from developing. A classic example is *RB1* which acts as a negative regulator of the cell cycle preventing uncontrolled cellular growth and malignancy[112]. Mutational inactivation of tumour suppressor genes, such as *RB1*, can be one of the initiating events that contributes to cancer development and acquisition of a number of the 'hallmarks of cancer'[7]. Oncogenes, in contrast, are activated by mutation. A classic

example of an oncogene is *MYC*. Many oncogenes like *MYC* are transcription factors so deregulation by mutation can lead to flow on effects to multiple genes[113]. *MYC* and a number of genes it interacts with are frequently deregulated in cancer. For example, the translocation t(8;14)(q24;q32), affecting *MYC*, is a hallmark of Burkitt lymphoma (BL)[63]. Together, mutations of tumour suppressor genes and oncogenes are the base upon which cancer develops. Mutations of tumour suppressor genes and oncogenes can be inherited, predisposing to cancer development, or acquired somatically as cancer develops, further driving malignancy.

There are now at least 114 clinically confirmed cancer predisposition genes, in a range of cancers, most of which have been identified through studies of families with unusual clustering of cancer occurrences[111]. Identification of genes in other familial cancers has facilitated the development of screening programs for relatives of cases. Through these programs relatives can assess their risk of developing disease allowing increased clinical monitoring and the option of prophylactic treatment, where applicable, for those relatives identified as having a greater genetic risk of cancer[114]. Importantly this approach has identified key targets for therapeutic research into these cancers. It is possible that genetic studies of family HMs will also identify cancer predisposition genes allowing development of HM screening programs and targeted therapies. A range of approaches has been taken in this area with a number of successful studies identifying HM predisposition genes. Figure 1.6, a Circos plot[115], summarises the findings of these varied approaches, discussed below.

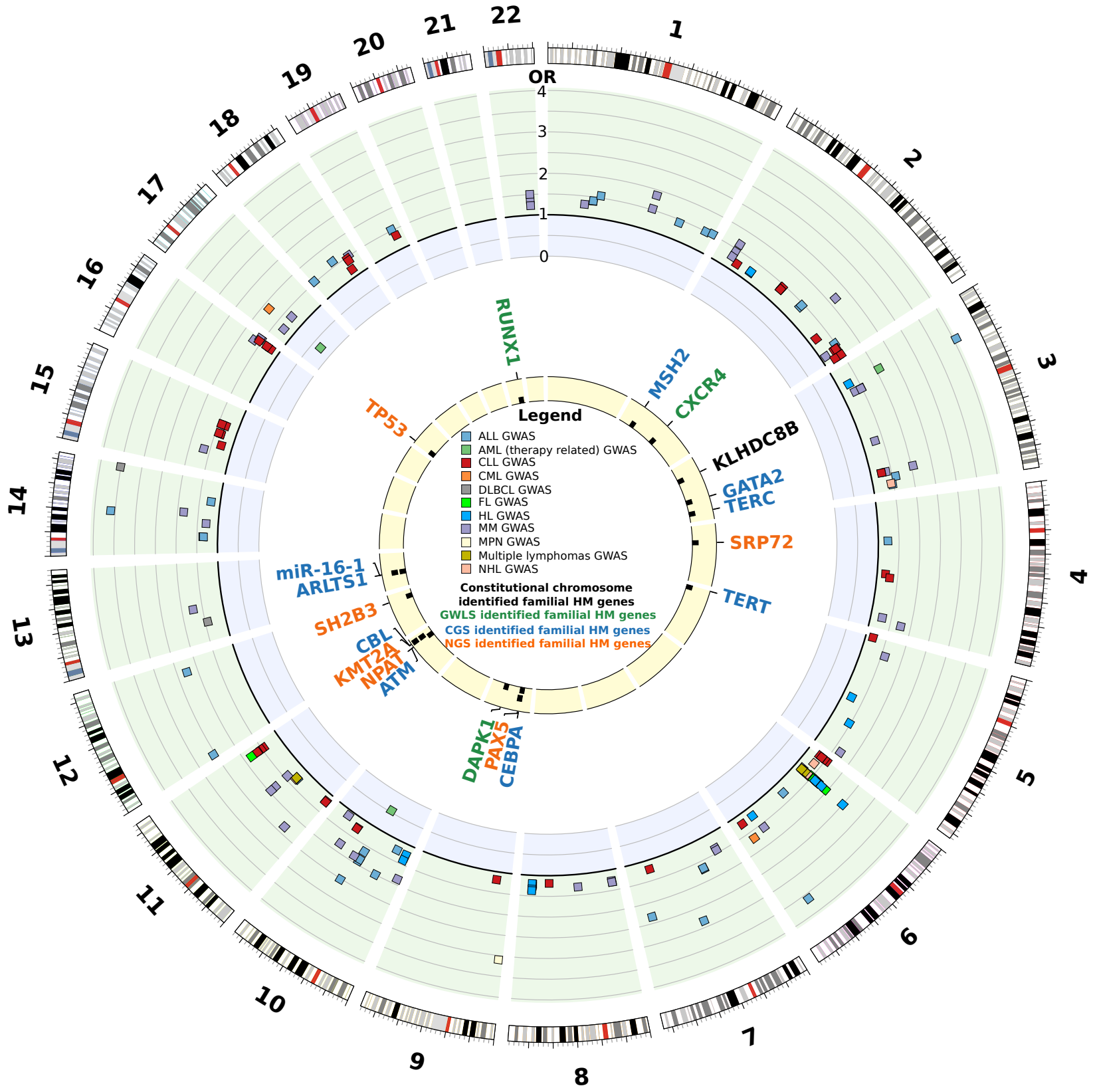


Figure 1.6 Circos plot graphical representation of currently implicated genes and loci in the germline susceptibility to HMs.

As detailed in Tables 1.5, 1.6, 1.9 and Appendix 1.1. Outermost track shows an ordered chromosome ideogram as per numbering; sex chromosomes are not shown, as there are no known associations. The second track shows a scatter plot of the odds ratios of the significant genome-wide association study (GWAS) findings, with green track background being odds ratios (ORs) > 1 and blue < 1, with OR = 1 shown by the thicker black axis, an axis scale is shown for this track at the 12'oclock position. The innermost track shows the known familial HM genes, labelled, identified through constitutional chromosome studies, genome-wide linkage studies, candidate gene studies and next generation sequencing studies.

1.5.1 Constitutional chromosome abnormalities in familial HMs

The first attempts to characterise the genetic factors contributing to familial HMs were cytogenetically-based. A study from 1969 identified a recurring abnormal karyotype present in both bone marrow and skin fibroblasts from a family with three CLL cases[116]. Other unaffected relatives also carried the abnormal karyotype but were unaffected by CLL. It is possible that carriers of the abnormal karyotype were at an increased lifetime risk of CLL, or that other genetic factors present in the three cases were contributing to disease. The specific genetic factors underlying the abnormal karyotype were not able to be determined with the technology at the time. Whether this abnormal karyotype contributes to HM development is unknown. A more recent study has investigated the occurrence of constitutional chromosome abnormalities in bone marrow from 5633 HM patients compared to non-disease tissue such as skin or remission samples[117]. Whilst this study was not family-based, it does show that 50 patients had constitutional chromosomal abnormalities that were germline or de novo in origin.

A recent study of a family with HL[118], identified a constitutional chromosomal breakpoint, involving chromosomes 2 and 3 and the disruption *KLHDC8B*. In this a family several individuals with the chromosome abnormality have also had HL. This study added further supportive evidence that constitutional chromosomal abnormalities may contribute to HM predisposition. *KLHDC8B* on chromosome 3 encodes the protein kelch domain containing 8B which is translated only during mitosis and has a role in cytokinesis[118]. Salipante and colleagues determined that a decrease in the expression of *KLHDC8B* during cytokinesis results in the binucleated Reed-Sternberg cells that are a pathological hallmark of classical HL. A study seeking to validate the findings of germline mutations in *KLHDC8B* in familial NLPHL did not find any disease associated mutations[119], which is consistent with Reed-Sternberg cells not being a feature of most NLPHL patient samples[24]. For *KLHDC8B* further studies are required to determine whether this gene contributes more widely to disease in other HM families.

1.5.2 Microsatellite and SNP-based genome-wide linkage studies of familial HMs

The development of Genome-Wide Linkage Studies (GWLS) using microsatellite repeat polymorphisms to identify genome linkage regions within disease families provided an increase to the resolution of genetic studies for detecting germline predispositions to disease. Researchers applied the GWLS technology to genome-wide investigations of families with multiple HM cases. Typically studies focused on families with one subtype of HM with GWLS conducted in families with CLL, HL, WM and AML. Table 1.5 summarises the findings from GWLS that have led to identification of genes contributing to familial HMs.

Table 1.5 Summary of genes identified by GWLS in familial HMs.

HM Subtype	Chromosome Band	Gene implicated	Method	Reference
FPD/AML	21q22.12	*RUNX1*	Microsatellite	[120,121]
CLL	9q21.33	*DAPK1*	Microsatellite	[122]
CLL	2q21.1	*CXCR4*	SNP	[123,124]

1.5.2.1 RUNX1 in familial platelet disorder with AML predisposition

The most influential finding from GWLS in HMs came from studying families with familial platelet disorder (FPD) and a predisposition to develop AML. This lead to the identification of germline mutations in *RUNX1*[120,121]. *RUNX1* encodes a transcription factor located on chromosome 21q22, that has a key role in haematopoiesis in the regulation of both myeloid and lymphoid lineages[8,125]. Additionally, Down syndrome or trisomy 21 has also been shown to have an increased predisposition to the development of HMs[23,126] for which the extra copy of *RUNX1* may be contributing. *RUNX1* mutations have since been consistently identified in FPD with predisposition to develop AML[127,128] and sporadic de novo AML, particularly as somatic mutations involving chromosomal translocations that create *RUNX1* gene fusion products such as *RUNX1-ETO*[129-132]. It is interesting to note here that the identification of germline mutations in *RUNX1* in a syndrome with a predisposition to develop AML has also been relevant in non-syndrome sporadic AML.

Current studies show that *RUNX1* is a tumour suppressor gene and that germline mutations act as a cancer initiating event, increasing the risk of AML, with further

somatic mutations required to fully develop AML[133]. There is also evidence implicating *RUNX1* in the development of ALL[133].

1.5.2.2 SNP array-based GWLS in familial CLL

With the development of SNP genotyping arrays, GWLS increased in resolution from hundreds of microsatellite markers to early SNP-based studies using around 10,000 polymorphic markers across the genome. This higher resolution of GWLS was applied in a number of studies of familial CLL. In 2005, Sellick and colleagues, using 115 CLL families found suggestive linkage at chromosome 11p11[134]. In an effort to increase the linkage detection power in their studies the two main groups studying familial CLL, one in the UK the other in the US, combined their familial resources for a SNP-based GWLS of 206 CLL families (155 CLL only and 51 families with CLL and other HMs), superseding all previous familial CLL GWLS[123]. Genome-wide significant linkage regions were detected on chromosome 5q23.2, 6p22.1, 11q12.1, 18q21.1 and 2q21.1. A follow up study of a SNP nearby the chromosome 18q21.1 region in 18q24 in 984 cases and 4831 controls did not replicate the association[135]. The 2q21.1 region was also followed up in a study that sequenced the *CXCR4* gene in this region in 188 familial CLL cases and 213 controls. The study identified three *CXCR4* germline mutations in CLL cases not present in controls, suggesting that rare *CXCR4* mutations may contribute to CLL susceptibility[124]. Despite these potential regions of interest in familial CLL, a GWLS of a single large multigenerational CLL family using SNP arrays did not find new or confirm previous linkage regions[136].

1.5.3 Genome-wide association studies of HM disease susceptibility

Even with the increase in resolution with SNP array technology, familial studies largely failed to find statistically significant evidence of linkage to loci associated with HM occurrence[136]. This may have been due to inadequate power from studies using only a few small families, or inadequate resolution with the small microsatellite marker numbers and the early lower resolution SNP arrays. With the increasing resolution of SNP-based arrays a new type of study emerged, the genome-wide association study (GWAS). A GWAS in concept is similar to a GWLS, looking for genetic variation more common across multiple cases than controls to implicate genomic regions in disease susceptibility. However a GWAS is fundamentally

different to a GWLS in that it's design is population-based instead of family-based, requiring large numbers of cases and controls for adequate power and specifically targets common variation, SNPs, contributing to disease susceptibility.

The first GWAS in HMs was conducted in a CLL cohort consisting of 517 CLL cases, 155 of whom had a family history of disease[137]. Initial associations identified were validated in two further cohorts identifying seven SNPs in six genomic regions associated with CLL risk. Associations were found with loci at 2p13, 2q37.1 (*SP140*), 6p25.3 (*IRF4*), 11q24.1, 15q23 and 19q13.32 (*PRKD2*) with odds ratios ranging from 1.39 to 1.46, providing the first evidence of low risk common variants predisposing to HMs. Since this first study, several other GWASs have been conducted, particularly in CLL but also other subtypes of HMs. Appendix 1.1 presents a summary of the loci identified to date associated with HMs, with significant associations and their odds ratios plotted on the radial axes of the Circos plot[115] in Figure 1.5.

1.5.3.1 CLL GWAS findings

CLL has been the most frequently examined HM using the GWAS approach. As at October 2014 there are 42 reported CLL GWAS associations with odds ratios ranging from 1.18 to 1.64 (Table 1.5) with estimations that the known associations account for around 19% of the familial risk of CLL[138]. Most of these CLL associations are to SNPs in intergenic, intronic or regulatory regions with as yet unknown effect on function. A coding variant has been reported as associated with CLL located in *MYNN* with an odds ratio of 1.26 (Appendix 1.1)[138], however the SNP has no clear functional effect on the *MYNN* gene.

Many of the CLL GWASs have suggested, in one syntax or another, that as the associated loci are common variants within the Caucasian population they must therefore contribute significantly to the risk of developing CLL with those having more variants also having a corresponding greater risk. This logic seems flawed. Even though CLL accounts for a large portion of HMs in western countries it is still a relatively rare disease in the general population. It is likely, given their frequency in the population, that the common variants identified as associated with CLL have a very small influence on disease risk. Indeed GWASs as a whole have been

disappointing as there has been little clinical impact achieved as a result of the GWAS identified association signals. For CLL rigorous studies have yet to be conducted which demonstrate the clinical relevance of the 42 associated loci.

It is likely that genes with germline variations identified as contributing to HM development will also be mutated somatically in HMs. Examining genes implicated by CLL GWAS loci in the 378 CLL cancer genomes contained within COSMIC reveals that genes at the CLL GWAS loci, with the exception of *POT1* with 12 samples (3.2%) containing mutations, are not recurrently somatically mutated in CLL[108-110]. It is however possible that these CLL GWAS loci, and those identified in future studies, may be used as variables in prediction models together with age, sex, ethnicity and family history of disease to determine the likelihood of an individual developing disease. Further studies are required to determine their clinical utility given that these are common mutations with small effect sizes that also occur in the general population.

1.5.3.2 ALL GWAS findings

Childhood ALL is another HM subtype that has been studied by many GWASs with 42 associations identified as potential disease susceptibility sites (Appendix 1.1). Many of these sites show associations across multiple ethnicities. This is best demonstrated by a study conducted in a childhood onset ALL cohort containing European, African, Asian and Native American genetic ancestries[139]. This study validated previous European ethnicity-based findings of associations at *IKZF1*, *PIP4K2A*, *ARID5B* and *CEBPE* across multiple ethnicities[139]. Of the genes linked to ALL susceptibility by GWAS two genes, *IKZF1* and *CDKN2A*, show evidence for recurrent somatic mutations in ALL in COSMIC[108-110].

Across all HM GWASs, studies of ALL have identified the largest odds ratio risks associated with disease, for example one SNP rs17079534 upstream of *MYRIP* has an OR of 4.07, however the biological significance of this association to ALL development has not been identified.

1.5.3.3 Criticism of HM GWAS findings

The underlying basis of GWASs is the 'common disease, common variant' hypothesis, whereby common diseases are due in part to common variants present in > 5% of the general population[140]. By virtue of their design GWASs will identify common variants that individually or in combination confer an increased risk of disease[140]. They are not designed to detect all genetic variation contributing to disease and are significantly biased specifically to common variation. The associated loci for HMs account for little of the heritability of disease, indicating that there are other important heritable factors contributing to disease risk. This remaining unexplained heritability has been termed the 'missing heritability' of disease and is repeatedly described for many complex diseases[140]. It has been suggested that the 'missing heritability' of complex diseases such as HMs may include genetic factors that are not detectable by GWAS such as rare germline variants, structural variants, crosstalk between genes, epigenetic factors, or interactions between genes and environmental factors[141].

It has been suggested that GWAS variants that have been identified across multiple diseases actually represent synthetic associations, particularly as GWAS variants are generally in genomic regions with no functional indication such as intronic and intergenic regions. First proposed in 2010 a synthetic association is when a rare variant or variants, that are the actual genetic factors contributing to disease susceptibility and not detectable by GWASs, occur by chance more often with one allele at a SNP, or sets of SNPs, than another thus driving the association at this loci[142]. Until recently synthetic associations only existed theoretically and have been a source of debate in the field[143]. However a recent study in prostate cancer has found that for the *HOXB* locus at 17q21 there is a synthetic association detected through common GWAS variants. However this is caused by an underlying partially correlated rare coding variant in *HOXB13*; the first identified synthetic association in cancer[144]. This is early evidence that supports the idea that the common GWAS variants associated with disease susceptibility in other diseases, such as HMs, may be due to synthetic associations. This does not make the findings of GWAS irrelevant but instead indicates genetic studies with greater resolution, such as is afforded by next generation sequencing, are the next step to identifying disease susceptibilities. In line with this there has been a shift in the GWASs of HMs away from attributing

disease susceptibility and causation to associations with common variants[137] towards using the results of GWASs to indicate genome localisations where the true susceptibility loci or candidate genes may be in close proximity[145].

1.5.4 Candidate gene approach to familial HMs

Genome-wide approaches, such as GWLSs, are an unbiased approach to identify HM predisposition genes. An alternate approach is the candidate gene approach which is a biased approach focusing on genomic regions or genes that have evidence implicating their involvement in disease. The candidate gene approach has resulted in identification of several familial HM genes, as summarised in Table 1.6. Particularly interesting is the identification of mutations in the telomere biology genes *TERT* and *TERC* and the haematopoietic transcription factor *GATA2* in familial MDS/AML.

Table 1.6 Summary of findings from candidate gene approaches to familial HMs.

HM Subtype	Gene	Reference
AML	*CEBPA*	146
MDS/AML	*TERT*	147
	TERC	147
	GATA2	148
CLL	*ARLTS1*	149
	miR-16-1	150
	ATM	151
T-cell NHL	*MSH2*	152
JMML	*CBL*	153

1.5.4.1 TERT and TERC mutations in familial MDS/AML

Germline mutations in *TERT*, the gene encoding telomerase, and *TERC*, the gene encoding the RNA component of telomerase, have been identified in 4 families with multiple cases of MDS/AML[147]. Telomeres are hexameric nucleotide repeats found at the ends of chromosomes. They are several kilobases in length and have an important role in protecting chromosomal DNA from the end replication problem during DNA replication. Telomeres and mutations in telomerase-associated genes have also been implicated as a mechanism of disease anticipation in the bone marrow failure disorder dyskeratosis congenita[154,155] and other diseases[156,157]. Further, patients with dyskeratosis congenita are at an increased risk of developing MDS/AML[88] and

germline mutations in *TERT* had also been identified in sporadic cases of AML[158] thus prompting the investigation of *TERT* and *TERC* in familial MDS/AML. In the four families described by Kirwan and colleagues, two families have germline mutations in *TERC* and two families have germline mutations in *TERT*. Not all mutation carriers in these families are affected with MDS/AML and there seems to be some diversity in the disease status of those carrying the mutations including the occurrence of aplastic anaemia. Given this diversity and the occurrence of similar mutations in dyskeratosis congenita it is likely that mutations in *TERT* and *TERC* predispose to a range of diseases with a strong predisposition to the development of MDS/AML. The mutations to *TERT* and *TERC* alter the activity levels of the telomerase protein, which in turn reduces the maintenance of telomeres, providing an opportunity for aberrant chromosomal interactions to occur and malignancies to develop. Germline mutations in *TERT* and *TERC* have yet to be confirmed as a recurring feature of familial HMs, however Kirwan and colleagues show evidence supporting a role for telomere biology in the development of HMs that warrants further investigation.

1.5.4.2 GATA2 mutations in familial MDS/AML

Together with *RUNX1*[120,121] and *CEBPA*[146], germline mutations in *GATA2* have been identified in affected individuals of families with MDS and AML[148]. By screening 50 candidate genes for variants, Hahn and colleagues found 4 families with novel germline *GATA2* variants. Three MDS/AML families were identified with the amino acid substitution Thr354Met and the fourth family consisting of a father-son pair, both affected by MDS, had a 3 base pair deletion in *GATA2* resulting in the amino acid deletion Thr355del. All family members affected with AML or MDS had a variant in *GATA2* but unaffected family members were also variant carriers. *GATA2* variants were not detected in 35 other families with HMs or 695 normal controls in this study. *GATA2* encodes a zinc-finger transcription factor located on chromosome 3. The two germline variants reported disrupt the zinc finger 2 domain of GATA2 resulting in a reduced ability for DNA binding and thus reduced transactivation ability on GATA2 responsive enhancers[148]. *GATA2* has been shown to have important roles in the myeloid cell pathway of haematopoiesis[159] which aids in our understanding of how mutations in *GATA2* contribute to myeloid malignancies including somatic mutations

in sporadic AML and CML[160,161]. The presence of *GATA2* variants in unaffected relatives of patients with HMs suggests that it is a predisposing germline variant not sufficient for malignancy development. Additionally, germline *GATA2* variants have also been identified in Emberger syndrome[162], MonoMAC syndrome[163] and an immunodeficiency syndrome characterised by a dendritic cell, monocyte, B and NK lymphoid deficiency[164] each of which predisposes patients towards the development of MDS and myeloid leukaemias. *GATA2* variants have been validated in other families with MDS/AML[165] as well as MDS/AML families that exhibit multiple *GATA2* related phenotypes including MonoMAC and primary lymphedema[166]. Identification of *GATA2* as a gene involved in familial HMs has contributed significantly to the understanding of how HMs develop but is yet to reveal new therapeutic avenues for treatment of HMs. The finding of multiple disease phenotypes resulting from *GATA2* variants has raised the possibility of a *GATA2* syndrome with one of the main features being a predisposition to HMs, dependent upon other genetic factors[167]. An example of one of these other genetic factors are somatic variants in *ASXL1* which are recurrently acquired in *GATA2* mutation carriers, possibly contributing to HM development[168].

1.5.4.3 Bias in candidate gene studies

In the context of familial cancers, including HMs, the use of candidate gene studies is inherently biased by a gene list selected by the investigators. Only mutations in genes in that list are considered, with varying degrees of evidence to support selection of those genes for study. Given that these studies screen for mutations in candidate genes in related cases it is not unreasonable to expect rare or even private mutations to occur by chance in at least one of the selected genes in the family. This means that results from biased approaches must be validated in multiple ways as they only provide the first indicative evidence that a gene is implicated in the familial risk of HMs. In the case of *CEBPA*, *ATM*, *TERT/TERC*, *MSH2*, *CBL*, and *GATA2* there is evidence from somatic studies of HMs and other cancers as well as functional studies that support the role of these genes in HM development. However, for the HM families that these genes were identified in, an additional unbiased analysis of their genomes may identify multiple other variants that may be involved in the familial predisposition to HM development.

HMs, and cancer more broadly, are not single gene diseases. Mutations in multiple genes will contribute to the familial predisposition. The degree of uncertainty from candidate gene approaches, and the resolution challenges with genome-wide linkage and association approaches are clear indicators that identification of familial HM predisposition genes will require a new approach. The advent of next generation sequencing (NGS) technologies such as whole genome sequencing (WGS) and whole exome sequencing (WES) presents a new and unbiased approach to identifying cancer predisposition genes in familial HMs.

1.5.5 HM predisposition as part of cancer predisposition syndromes

Within a number of inherited cancer predisposition syndromes such as Li Fraumeni syndrome there is an increased risk of HM development[169]. HM families have been described carrying mutations in genes which are more typically implicated in a cancer predisposition syndrome (such as *TP53* discussed in section 1.10.3.3) thus there is an overlap between the genes identified as causal in both HMs connected to a syndrome and familial HMs without a known underlying syndrome. Rahman[111] has curated and reported a set of 114 cancer predisposition genes, many of which are associated with a cancer syndrome. Of the genes reported by Rahman, 19 have a HM as a major associated tumour type, as shown in Table 1.7. It is important to be aware of these syndrome associated HM genes as evidence suggests that the genes predisposing to HMs associated with cancer predisposition syndromes may also be implicated in non-syndrome familial HMs, as has been seen for *TERT*/*TERC*[147] and *GATA2*[148].

Table 1.7 Cancer predisposition genes with HM as major associated subtype, adapted from Rahman[111].

Gene	Syndrome/s	HM subtypes as reported by Rahman
ATM	Ataxia-telangiectasia	Lymphoid HMs, lymphoma
BLM	Bloom syndrome	Lymphoma, ALL, myeloid HMs
BRCA2	Hereditary breast-ovarian cancer syndrome and Fanconi anaemia (D1)	Myeloid HMs
BRIP1	Fanconi anaemia (J)	Myeloid HMs
BUB1B	Mosaic variegated aneuploidy syndrome	Myeloid HMs
CBL	Noonan syndrome	JMML
DKC1	Dyskeratosis congenita	AML
DOCK8	Hyper-immunoglobulin E syndrome	Lymphoma
ELANE	Severe congenital neutropenia	Leukaemia
FANCA	Fanconi anaemia (A)	Myeloid HMs
FANCC	Fanconi anaemia (C)	Myeloid HMs
FANCG	Fanconi anaemia (G)	Myeloid HMs
GATA2	Emberger MonoMAC syndrome	Myeloid HMs
GBA	Gauchers type 1	Myeloma, myeloid HMs
ITK	Lymphoproliferative syndrome 1	Hodgkin lymphoma
MLH1	MMR deficiency syndrome (biallelic mutations) Lynch syndrome / Hereditary non-polyposis colon cancer (monoallelic mutations)	HMs
MSH2	MMR deficiency syndrome (biallelic mutations) Lynch syndrome / Hereditary non-polyposis colon cancer (monoallelic mutations)	HMs
MSH6	MMR deficiency syndrome (biallelic mutations) Lynch syndrome / Hereditary non-polyposis colon cancer (monoallelic mutations)	HMs
NBN	Nijmegen breakage syndrome	Lymphoma
PALB2	Fanconi anaemia (N)	Myeloid HMs
PMS2	MMR deficiency syndrome (biallelic mutations) Lynch syndrome / Hereditary non-polyposis colon cancer (monoallelic mutations)	HMs
PTPN11	Noonan syndrome	JMML
RMRP	Cartilage-hair hypoplasia syndrome	NHL, leukaemia
RUNX1	Familial platelet disorder	Myeloid HMs
SBDS	Schwachman-Diamond syndrome	Myeloid HMs
SH2D1A	Lymphoproliferative disease	Lymphoma
STAT3	Hyper-immunoglobulin E syndrome	Lymphoma
TERT	Dyskeratosis congenitia	AML
TNFRSF6	Autoimmune lymphoproliferative syndrome	Lymphoma
WAS	Wiskott-Aldrich syndrome	Lymphoma

1.6 Next generation sequencing in HMs

Next generation sequencing (NGS) includes the technologies of whole genome sequencing (WGS), whole exome sequencing (WES) and RNA sequencing for whole-transcriptome sequencing. NGS affords a significant advancement upon previous Sanger sequencing capabilities with regard to throughput, cost and applications. NGS has facilitated a number of new approaches to study familial diseases such as HMs. It is now possible to obtain full coverage of the genome for multiple affected and unaffected members of a HM family and use this information to identify inherited mutations that may be contributing to disease.

Another NGS approach that has been applied to HM genetics is the paired tumour-normal sequencing approach where the tumour genome and germline genome for an individual are both sequenced from appropriate source tissues. In HMs this is typically DNA from circulating blood malignant cells or solid-tumour biopsies for the tumour genome and DNA from a skin biopsy for the germline genome. The variants identified in the germline genome are filtered from the tumour genome leaving only somatic mutations. Both the familial study and 'tumour-normal' study approaches are revealing important new insights into the genetics of HMs and both have considerable potential to have a clinical impact on patient treatment and diagnosis. The primary focus of NGS in HMs to date has been to identify somatic mutations with much less focus given to germline mutations.

1.6.1 Identifying recurrent somatic mutations in HMs using NGS

The paired tumour-normal approach has been successful in revealing a number of recurrently mutated genes across specific HM subtypes with significant clinical impact and large potential for targeted disease treatment. Indeed the first WGS report of a human cancer was the tumour-normal sequencing of a single AML patient identifying ten genes with somatic mutations of which eight were previously undescribed in HMs[170]. Since this first report in 2008 many other HM subtypes have been studied using WGS or WES identifying a range of genes that are frequently mutated, with some common patterns between HM subtypes[171], and even between HMs and other cancers[172]. What has become clear from this type of study is that the tumour profile for each HM diagnosed is unique but when large enough numbers of

tumours are sequenced mutation patterns are revealed with some subtype specific and cross-subtype recurring gene and pathway mutations. Table 1.8, adapted from[171] summarizes the major somatic mutation findings of WGS/WES studies in HMs. There are over 20 published genome and exome sequencing studies of somatic mutations in HMs. Contained within the whole genome section of the COSMIC is information for over 1003 HM tumour genomes in 19 different HM subtypes[108-110]. The most common recurring gene mutated is *TP53* in 9% of genomes, followed by *NPM1* in 5% of genomes[108-110], with many more genes mutated in lower sample percentages.

Table 1.8 Major findings from WGS/WES studies of HMs, adapted from Watson *et al.*[171].

HM subtype	Method	Sample size	Highlighted or novel somatically mutated genes*	References
AML	WGS	24	*TP53*, *KRAS*, *SMAD4*, *MLL3*, *ROBO2*, *RNF43*, *PEG3*, *GNAS*	173
	WGS	50	*SMC3*, *SMC1A*, *STAG2*, *RAD21*	174
	WES	150	*FLT3*, *NPM1*, *DNMT3A*, *IDH1*, *IDH2*, *TET2*, *RUNX1*, *TP53*, *NRAS*, *CEBPA*, *WT1*, *PTPN11*, *KIT*, *U2AF1**, *KRAS*, *SMC1A*, *SMC3*, *PHF6*, *STAG2*, *RAD21*, *FAM5C*, *EZH2*, *HNRNPK*	174
MDS	WES	29	*SF3B1**, *SRSF2*, *U2AF1**, *ZRSR2*, *SF3A1*, *PRPF40B*, *U2AF2*, *SF1*	175
	WES	9	*SF3B1**	176
CLL	WES	5	*NOTCH1*	177
	WGS	4	*NOTCH1*, *MYD88*, *XPO1*, *KLHL6*	178
	WES / WGS	88	*SF3B1**	179
	WES	105	*SF3B1**	180
DLBCL	WES	6	*MLL2*, *CREBBP*, *EP300*	181,182
	WES / WGS	13	*MLL2*, *MEF2B*	183
	WES	55	*MEF2B*, *MLL2*, *BTG1*, *GNA13*, *ACTB*, *P2RY8*, *PCLO*, *TNFRSF14*, *BCL2*	184
MM	WES / WGS	38	*DIS3*, *FAM46C*, *LRRK2*, *BRAF*, *IRF4*, *BTRC*, *CARD11*, *CYLD*, *IKBIP*, *IKBKB*, *MAP3K1*, *MAP3K14*, *RIPK4*, *TLR4*, *TNFRSF1A*, *TRAF3*	185

* *indicates that genes are mutated across multiple HMs.*

1.6.2 Application of NGS to familial HMs and germline mutations

Despite the repeated identification of a family history of disease as a risk factor for HMs and reports of individual families with multiple HM cases the NGS approach to identify germline susceptibilities has not been widely applied in HMs. Familial NGS studies involve the selection and sequencing of families with multiple related affected family members. Analysis then focuses on identifying shared mutations between related affected family members which are ideally not present in unaffected family members. In cancers such as HMs and other complex diseases, penetrance and susceptibility issues further complicate the study design. Not all mutation carriers develop a HM, nor will they develop HMs in the future.

The primary benefit of conducting a familial HM study over a population-based one is that studying families increases the chance of detecting rare mutations contributing to disease[186]. In a population study of cases it is likely that each case has different genetic factors contributing to their disease development and for each rare susceptibility mutation there will be only one copy present in the population. In a family study cases have a shared genetic background, which increases the likelihood of finding multiple copies of the disease susceptibility mutations. These mutations may be extremely rare in the population with minor allele frequencies less than 1%, or private mutations present only in the study family.

To date a small number of familial HM studies have used the NGS approach to identify susceptibility genes through exome sequencing in related cases and follow up in the rest of the family. This approach has identified six genes with germline mutations in specific HM subtypes as summarised in Table 1.9. Three of these genes, *TP53*[187], *SH2B3*[188] and *PAX5*[189] were identified in familial ALL studies. While mutations in the remaining three genes, *SRP72*, *NPAT* and *KMT2A*, were identified in family studies of MDS[190], nodular lymphocyte predominant Hodgkin lymphoma (NLPHL)[191] and primary mediastinal large B-cell lymphoma (PMBCL)[192] respectively. Of particular interest is the *PAX5* ALL study which identified the same mutation in both a Puerto Rican family and an African-American family[189]. The same gene has also been identified as mutated in an Ashkenazi Jewish family with ALL[193], highlighting that mutations in a single gene can cause HMs across three different ethnicity groups. *PAX5* is a transcription factor that has a key role in haematopoiesis,

specifically B-cell development, and has been shown to have diverse roles in malignancy development, as reviewed in[194]. It is known to be a recurrent acquired mutation in ALL[108,109,195,196] and in these families it appears that loss or inactivation of the second copy of *PAX5* is required for ALL to develop. The identification that an inherited *PAX5* mutation contributes to ALL in these families means that unaffected mutation carriers can undergo regular cancer surveillance to detect the early onset of ALL. This means that ALL in *PAX5* mutation carriers can be treated early and potentially with therapies specifically targeted at *PAX5* related haematopoietic differentiation pathways. Additionally other families with clustering of ALL can be screened for mutations in *PAX5* informing therapy options when mutations are identified, particularly in regards to selection of related bone marrow donors without the mutation as has occurred in another NGS study of familial ALL for *TP53*[187].

Table 1.9 Summary of findings from NGS approaches to familial HMs.

HM Subtype	Gene	Reference
ALL	*TP53*	[187]
	SH2B3	[188]
	PAX5	[189]
MDS	*SRP72*	[190]
NLPHL	*NPAT*	[191]
PMBCL	*KMT2A*	[192]

1.7 Family studies are most likely to identify HM genetic susceptibilities

The genetic architecture of HM susceptibility has yet to be fully characterised. A range of different methods ranging from candidate gene studies, to GWAS and now, NGS, has identified several susceptibility genes. With the power and resolution provided by NGS, GWAS and candidate gene studies while potentially still informative have been superseded. As discussed, there are two very clear branches of previous work in the field of HM genetics using NGS. One branch has focused on studies of somatic or acquired cancer mutations while the second branch has focused on studies of germline or inherited cancer susceptibility mutations. In the context of familial studies somatic mutations in cancer are of secondary interest to the primary study goal of germline mutations. The focus of this dissertation is the latter but it must also be recognised that one can inform the other; germline genetic factors in familial HMs may recapitulate the somatic cancer mutations of HMs and vice versa. Ultimately the successful treatment of HMs will lie in a combined assessment of the patient's germline genetic background, which could contain targetable factors driving tumour development, as well as an assessment of the complexity of their somatic tumour genome. Together these will provide the necessary information crucial for targeted genetic therapies and personalised medicine.

Identification of germline susceptibility mutations in HM families will enhance our understanding of HM development by describing the early genetic factors that contribute to the progression to malignancy. This will indicate pathways that can be targeted early in disease development for effective treatment before malignancy and multi-clonality of disease is established. It will provide the opportunity to screen individuals, particularly those with a family history of disease, for susceptibility mutations. A positive screening result can then lead the individual into a cancer surveillance strategy to catch any development of HMs early, which depending upon the subtype may inform a targeted treatment regimen with an increase in positive treatment outcomes. Finally, for HM families, identification of germline mutations will provide opportunities for genetic counselling of HM risk and informed lifestyle choices.

It is also important to consider the risks associated with identifying new genetic susceptibilities underling predispositions to HMs, or indeed other types of cancer. While having significant research application and long-term clinical potential, there may be of no immediate clinical utility for families identified to carry the mutation/s. Indeed the identified germline mutations may be only one part of a larger genetic profile requiring multiple additional somatic genetic changes to occur for cancer to develop. Further, screening of unaffected relatives of familial HM cases in a genetic counselling context, when no preventative treatments are available, could unnecessarily increase anxiety and stress within the family. It has been suggested that population based cancer screening strategies are not resulting in the expected improved health outcomes from early disease detection. Evidence supporting this is derived from population-based screening studies conducted in ovarian cancer[197] and breast cancer[198]. For those individuals with a known family history of cancer, the benefits of screening include both the management of anxiety arising in those concerned about a history of disease in their families, as well as the possible improved health outcomes from early cancer detection. Population screening and early detection of the onset of HMs is currently of limited clinical value, however there may be clinical relevance in families with an increased risk of disease. Encouragingly in other cancers, at least for carriers of mutations predisposing to Lynch syndrome (i.e. individuals with a family history of disease) there is a demonstrated benefit for decreasing colon cancer occurrence by using aspirin as a chemoprevention treatment[199].

The short-term benefits and risks of identifying HM predisposing mutations should not be overstated. Instead the focus of research in this area, after mutation identification, should be the potential long-term clinical benefits for identifying a recurring mutation, or mutated gene that affects multiple families, with demonstrable disease causality, for which targeted therapies can be generated.

1.8 Hypothesis and aims of the study

This study explores the area of the genetic susceptibility of HMs using HM families. To date, mutations in several genes have been identified that contribute to the genetic predisposition to HMs (Figure 1.6). Many of these genes are also somatically mutated in HM tumour genomes. Known genes account for only a small portion of the overall

inherited component of HMs leaving a significant gap in our understanding of the genetic basis of disease. Application of innovative NGS technologies presents an ideal opportunity to generate new findings in this area. To date, familial HM studies have generally focused on families with single subtypes of HMs while evidence shows that in relatives of HM patients there is an increased risk of multiple HM subtypes (see Table 1.4, Figure 1.5). Therefore this study hypothesises that:

Studies of families with multiple HM subtypes will identify genetic factors contributing to disease predisposition.

Two specific aims were developed to test this hypothesis:

Aim 1: Explore the use of a NGS approach in the genetic analysis of familial HMs, using extended pedigrees from the Tasmanian population, to identify candidate variants that are likely to predispose to disease. The familial dataset used for this study, including pedigree structure and clinical information for families that were sequenced, is described in Chapter 2. Chapter 3 details the NGS assembly and variant calling pipeline developed and used, and describes the prioritisation strategies used to identify variants for the laboratory-based follow-up analyses in Chapter 4.

Aim 2: Use data from a quantitative trait, telomere length, due to its relevance to cancer development, in variance components modelling to identify whether variation in telomere length as a trait is related to HM status in this resource. These analyses are described in Chapter 5.

Chapter 6 draws together the findings of this dissertation linking them to the broader clinical sphere for HMs and makes recommendations as to the future use of susceptibility information in clinical practice for treatment of HMs.

Chapter 2 - The Tasmanian Familial Haematological Malignancies Study

2.1 Background

2.1.1 Genetic studies in Tasmania

Tasmania is the island state of Australia and had a population of over 513,400 at September 2013[200]. Tasmanians are predominantly of Caucasian descent with an estimated 87% of Tasmanian residents born in Australia[201], the highest of all Australian states and territories. In the 2011 Australian Census the top ancestries reported by Tasmanians were Australian (33.9% responders), English (33.7%), Irish (7.8%) and Scottish (6.7%)[202].

Indigenous people first inhabited Tasmania prior to European contact with an estimated population of 5000-10000. The British colonised Van Diemen's Land, as Tasmania was then known, in 1803. Initially Tasmania was established as a penal colony, with a small number of free settlers and military personnel sent to develop an agricultural industry. Convict transportation to Tasmania ceased in 1853 with around 75,000 convicts transported[203]. With Australian federation the colony of Tasmania became the state of Tasmania in 1901. Over this time the population grew to become a mixture of immigrating British free settlers and released convicts[203].

The result of this is that Tasmania is unusual in comparison to other states of Australia due to its largely homogenous population of residents with British ancestry. An estimated 65% can trace their heritage back to around 10000 founder families from the mid-19th century. This together with Tasmania's historically low migrant rates created a closed population with little movement between Tasmania and the Australian mainland. Today, of course, immigration into and out of Tasmania is much higher but overall Tasmania still has the lowest migration rates in Australia[201].

The early history of the Tasmanian population, including convict transportation, free-settler land ownership, and early British immigration, is well documented. This together with the establishment in 1838 of an Act requiring civil registration of all

births, deaths and marriages, means early generations of the Tasmanian population are very accessible for genealogical research. This facilitates the identification of large multigenerational families with many individuals from the current generations still living in Tasmania.

Such genealogical potential has facilitated past and present familial genetics research programs in Tasmania. Indeed for a number of conditions a founder effect has been observed in specific Tasmanian families. An early prominent example of this includes a large Tasmanian family with Huntington's disease. All Huntington's disease in Tasmania can be traced back to a female immigrant who migrated to Tasmania in 1843 with her husband and seven children[204]. Another example is the 'Tasman 1' multiple endocrine neoplasia (MEN1) family that can be traced back to a single English immigrant and his descendants, of which there were around 2000 in the original publication, with other independent MEN1 families identified in Tasmania[205]. Importantly, the genealogical knowledge of these families has allowed the tracing of disease inheritance through the generations.

Huntington's disease and multiple endocrine neoplasia are both examples of single-gene 'Mendelian' diseases with mutations in the *HTT* and *MEN1* genes respectively. Tasmania also has a history of familial studies of complex, multi-genic diseases such as prostate cancer[206], primary open angled glaucoma[207], and multiple sclerosis[208]. Another complex familial disease studied in Tasmania is haematological malignancies, on which this dissertation is based[92,97,209-212].

2.1.2 The Tasmanian Familial Haematological Malignancies Study

The Tasmanian Familial Haematological Malignancies Study (TFHMS) is a collection of large Tasmanian families with multiple generations affected by a range of HMs. The TFHMS owes its origins to a population-based study of HMs in Tasmania conducted by the University of Tasmania and the Royal Hobart Hospital between 1972 and 1980[209-211]. The original study identified some occupational and demographic exposures as disease risk factors. For example farmers, miners and hairdressers as well as people who lived in rural areas were at a heightened risk of

disease. An additional finding was the clustering of disease in small family units which was at that time attributed to shared environmental factors[209,211].

Researchers involved in the earlier study, particularly the research assistant Jean Panton, observed that a number of the HM patients presenting for treatment were related and accordingly began to informally collect together large pedigrees with clusters of HM cases. In 2006 this resource became part of the Menzies Institute for Medical Research and formally became the Tasmanian Familial Haematological Malignancies Study. Using a genealogical database at the Menzies Institute for Medical Research the original 866 participants from the population-based study were linked to both current generations and disease records from the Tasmanian Cancer Registry, which has documented cases of HMs since 1978. A public recruitment campaign was conducted and many of the original study families as well as new families volunteered or were invited to join the TFHMS. Family members provided further information through questionnaires and personal interviews. The result is a statewide resource of multigenerational families with multiple cases of HMs as well as a collection of HM cases with no reported family history of disease. Historical information from clinical records, patient self-report and Jean Panton's records, was available in some instances allowing data for deceased cases to be included.

Previously the TFHMS resource has been used to identify whether particular subtypes of HMs cluster together in families[92]. Analyses showed that CLL cases in particular clustered together in families but did not cluster with other B-cell malignancies. Additionally evidence suggesting a shared genetic predisposition to both lymphoid, and myeloid malignancies, was found on the basis that families were not enriched solely for a single type of HM. Previous research using this resource also identified evidence for the genetic phenomenon of anticipation whereby the age of disease diagnosis is earlier with each subsequent affected generation in the family[97]. These findings strengthen the argument that in these families there is a genetic basis to HM occurrence.

2.1.3 Definition of affection status

In this study HM cases are defined according to the 2008 World Health Organisation classification system for tumours of haematopoietic and lymphoid tissues[24]. Initial diagnoses were obtained from reports from the Tasmanian Cancer Registry. For cases predating the registry (pre-1978) diagnosis was obtained from family records, death certificates, or archived medical records and where possible diagnostic criteria were matched to current WHO guidelines. For more recent and living cases for whom we have DNA samples, diagnosis was confirmed where possible by review of patient medical records by clinical pathologists (Dr. Katherine Marsden and Dr. Elizabeth Tegg) and a clinical oncologist (Prof. Ray Lowenthal).

2.2 Methods

2.2.1 Ethics approval

Ethics approval for this study was obtained from the Human Research Ethics Committee Tasmania Network and has been continuously in place since its inception in 2005 (HREC reference number H8551). This study was conducted in accordance with the Australian National Statement on Ethical Conduct in Human Research 2007 (updated March 2014)[213], and the Australian Code for the Responsible Conduct of Research[214]. Written informed consent was obtained from all study participants.

2.2.2 The TFHMS Resource

The TFHMS resource consists of 48 multigenerational families with diverse subtypes of HMs from whom we have obtained DNA samples, where possible, from HM cases and their unaffected relatives, predominantly first-degree relatives. We also have a collection of 84 non-familial HM cases that have no self-reported family history of disease and after extensive genealogical evaluation are not part of any known Tasmanian HM families. Table 2.1 summarises this resource. Families selected for intensive analysis in this dissertation include LK0051, LK0139, LK0124, LK0153 and LK2042 as indicated in Table 2.1.

Table 2.1 Summary of TFHMS Families.

Family	Known HM cases	Generations with HM cases	HM Cases with DNA	Unaffected relatives with DNA
LK0001	14	4	4	18
LK0002	15	3	2	12
LK0004	7	2	3	11
LK0016	18	5	10	31
LK0024	3	2	1	0
LK0026	6	2	1	6
LK0040	7	4	2	2
LK0051*	**21**	**5**	**9**	**29**
LK0054	9	3	0	4
LK0065	8	2	1	8
LK0124*	**24**	**5**	**13**	**40**
LK0132	5	2	0	8
LK0139*	**7**	**2**	**2**	**2**
LK0153*	**9**	**2**	**4**	**8**
LK0511	2	2	1	0
LK0512	2	1	1	0
LK0537	2	1	2	0
LK0546	2	2	1	0
LK0560	2	2	1	0
LK0561	2	2	1	0
LK0580	2	2	2	0
LK0600	5	3	3	2
LK0625	4	2	2	1
LK0647	2	2	1	0
LK0672	3	3	1	0
LK0710	3	2	0	1
LK0823	2	1	2	0
LK0836	6	3	2	6
LK1155	2	1	1	3
LK2042*	**32**	**5**	**15**	**48**
LK2447	3	2	1	2
LK6000	6	2	3	0
LK7739	2	1	1	0
LK7740	2	2	2	0
LK7743	3	2	2	0
LK7744	2	2	0	1
LK7748	2	2	1	0
LK7749	3	2	1	0
LK7750	4	2	2	0
LK7751	9	3	1	0
LK7754	3	1	1	0
LK7755	2	2	2	2
LK7756	2	1	1	1
LK7766	2	2	1	0
LK7768	2	1	1	0
LK7772	4	2	1	0
LK7773	2	1	2	0
Total	-	-	112	246
Non-familial cases	-	-	84	1

*Indicates families selected for analysis by NGS in this dissertation

2.2.3 Population controls

Tasmanian population controls used in this study were recruited randomly from the Tasmanian electoral role (n = 758), through the Tasmanian Study of Cognition and Gait (TASCOG) study (HREC reference number H9785)[215] and from population controls in the prostate cancer familial (HREC reference number H6914) and case-control (HREC reference number H7740) studies[206]. Table 2.2 summarises the characteristics of this population control set. In this study, different subsets of the population controls were used depending upon the analysis conducted.

Table 2.2 Summary of population controls.

Control Group	Number	Mean age (range)	% Male
TASCOG	393	73.2 (61.4 - 87.9)	57.5%
Case-control prostate cancer study controls	232	61.0 (45.1 - 71.3)	100%
Familial prostate cancer controls	133	61.9 (30.7 - 80.9)	90.2%
All	758	67.5 (30.7 - 87.9)	76.3%

2.2.4 Genetic material

For participants in the TFHMS genetic material for DNA was obtained primarily from peripheral blood samples but also from saliva samples using Oragene collection kits. Similarly for the population controls, DNA was obtained from peripheral blood samples.

HM patients who received an allogeneic bone marrow transplant prior to recruitment into the study or HM patients who may have circulating tumour cells within the sample obtained were noted. For these patients an Oragene saliva DNA sample was obtained, where possible, as an alternative, recognising that saliva DNA is known to be a mixture of both epithelial and white blood cells.

DNA from peripheral blood samples was extracted using Nucleon BACC3 Genomic DNA Extraction Kits (GE Healthcare Life Sciences) according to standard kit protocols. The Nucleon BACC3 system is a resin-based method of DNA extraction.

Saliva samples for DNA collected using Oragene 2 mL collection kits (DNA Genotek) were extracted according to standard kit protocols.

2.2.5 HM patient clinical information

Clinical Pathologist Dr Katherine Marsden and Clinical Oncologist Prof. Ray Lowenthal conducted a detailed clinical review of available medical records for HM cases in the TFHMS. Where available, clinical information included details of patient treatment and disease course, pathology reports and clinical notes. This meant that confirmation of HM diagnosis was possible and facilitated an assessment of whether patients had circulating disease at the time of blood sample collection.

2.3 Description of TFHMS extended pedigrees used in this study

2.3.1 Section overview

This section describes the TFHMS extended pedigrees and describes the available clinical information for those cases selected in these families for whole genome sequencing or whole exome sequencing. Each family is accompanied by a pedigree figure containing only the relevant parts of the family. From 1978 onwards, confirmation of case status is available for all participants. The disease status for earlier generations is generally unknown unless this information was obtained from clinical records and these individuals have been marked as unaffected in pedigree diagrams. Figure 2.1 describes the generic symbol representations used across the family pedigree drawings with pedigree specific information described on the relevant figures.

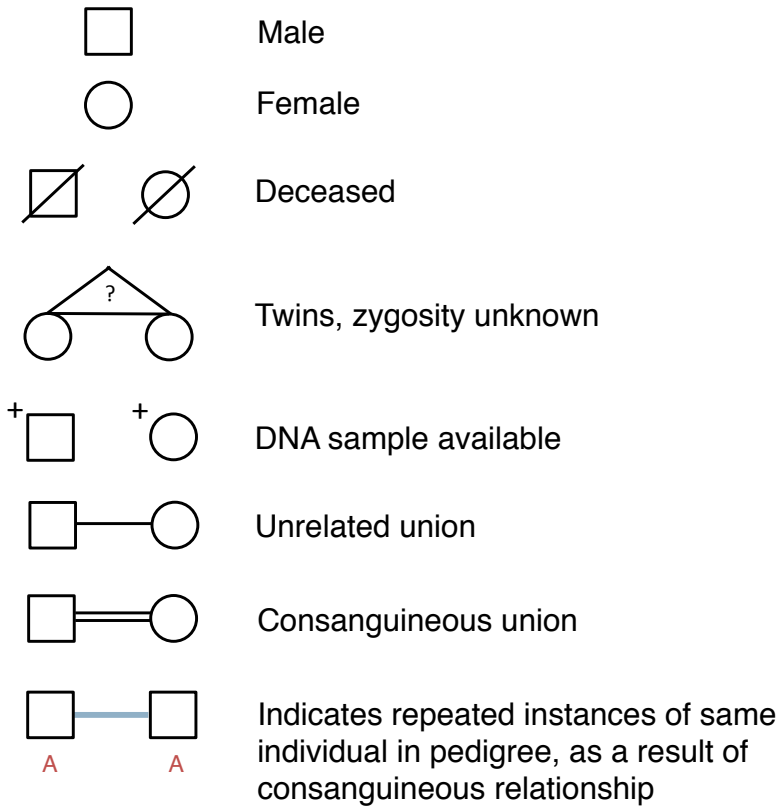


Figure 2.1 Generic pedigree symbol key explaining the symbols used throughout pedigree figures.

2.3.2 Family LK0051

LK0051 (shown in Figure 2.2) is a family affected by HMs in five generations. Nineteen HM cases have been identified along with one other cancer. There are seven different HM case subtypes in this family. Members of this family were ascertained from the original 1970s study[209-211]. Five people in this family were sequenced including three WGS cases and two WES unaffected relatives, as shown in Table 2.3 and Figure 2.2.

Table 2.3 Characteristics of sequenced family members in LK0051.

Sample ID	HM Diagnosis	Sex	Age at diagnosis	Age at sample collection	Sample types available*	Sample type for NGS	Evidence of circulating disease in sample	WGS or WES
LK0051-001	T-cell NHL	M	30	40	PB / O	PB	No	WGS
LK0051-007	Unaffected	F	--	46	PB	PB	--	WES
LK0051-128	BL	M	6	13	PB	PB	No	WGS
LK0051-159	DLBCL	F	58	69	PB / O	PB	No	WGS
LK0051-165	Unaffected	M	--	65	PB	PB	--	WES

* PB = genetic sample available from peripheral blood, O = genetic sample available from saliva collected in an Oragene kit.

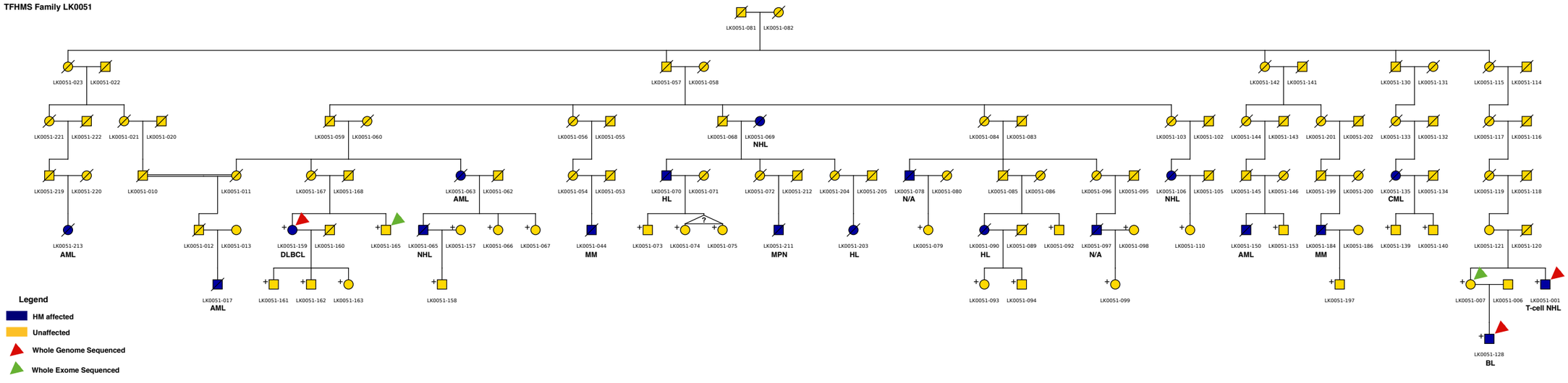


Figure 2.2 Extended pedigree of TFHMS family LK0051.
For the current study three HM cases were genome sequenced (LK0051-001, LK0051-128 and LK0051-159) and two unaffected relatives were exome sequenced, one as an informative control (LK0051-165) and the other as an obligate carrier (LK0051-007). Two of the genome sequenced cases (LK0051-001 and LK0051-128) have an avuncular relationship, connected by their mother/sister (LK0051-007). The remaining genome sequenced case (LK0051-159) and her unaffected exome sequenced brother (LK0051-165) are distant relatives to this trio, separated by nine meioses. Diagnosis abbreviations: AML = acute myeloid leukaemia, BL = Burkitt lymphoma, CML = chronic myeloid leukaemia, DLBCL = diffuse large B-cell lymphoma, HL = Hodgkin lymphoma, MM = multiple myeloma, MPN = myeloproliferative neoplasm, NHL = non-Hodgkin lymphoma, N/A = specific diagnosis unavailable.

2.3.3 Family LK0124

LK0124 (shown in Figure 2.3) is a large HM family with five generations affected by HMs. Members of this family were ascertained from the original 1970s study[209-211]. A total of 24 cases with several different subtypes of HMs occur in this family. This family has five consanguineous loops as a result of five cousin-cousin marriages. An unusual feature of this family, noted by haematologist Dr. Elizabeth Tegg, is that there are seven known cases of DLBCL including multiple atypical site presentations such CNS presentations in LK0124-179 (spinal presentation) and LK0124-187 (brain presentation). Three HM cases were genome sequenced in this family, as shown in Table 2.4 and Figure 2.3.

Table 2.4 Characteristics of sequenced family members in LK0124.

Sample ID	HM Diagnosis	Sex	Age at diagnosis	Age at sample collection	Sample types available*	Sample type for NGS	Evidence of circulating disease in sample	WGS or WES
LK0124-117	AML	M	24	36	PB / O	PB	No	WGS
LK0124-179	DLBCL (CNS)	M	84	87	PB	PB	No	WGS
LK0124-202	APML	M	30	35	PB / O	PB	No	WGS

* PB = genetic sample available from peripheral blood, O = genetic sample available from saliva collected in an Oragene kit.

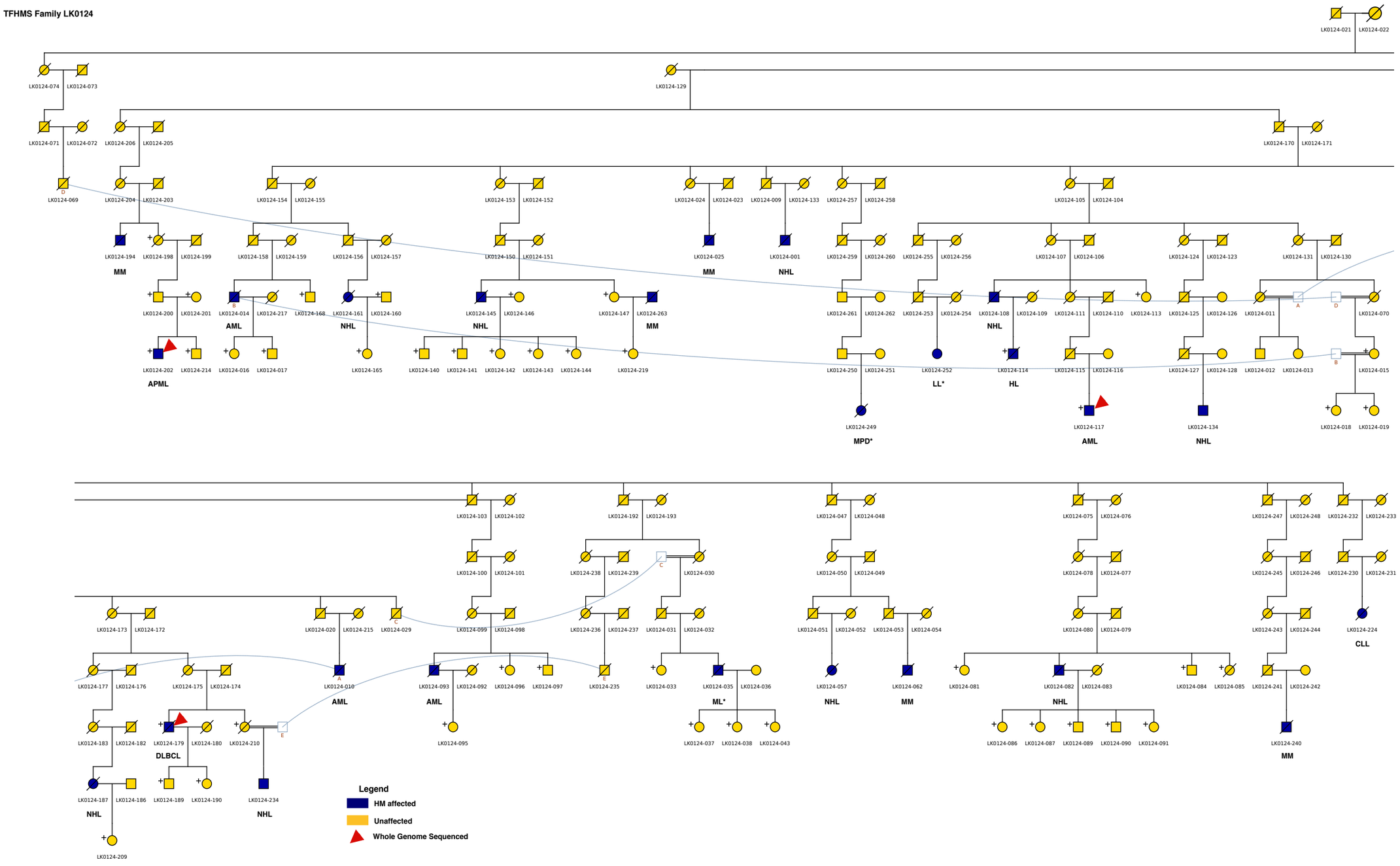


Figure 2.3 Extended pedigree of TFHMS family LK0124.
The focus in this family for the current study has been on three HM cases, LK0124-117, LK0124-202 and LK0124-179. These three cases are distant cousins. All were genome sequenced. Diagnosis abbreviations: APML = acute promyelocytic leukaemia, AML = acute myeloid leukaemia, CLL = chronic lymphocytic leukaemia, DLBCL = diffuse large B-cell lymphoma, HL = Hodgkin lymphoma, MM = multiple myeloma, ML* = myeloid leukaemia specific subtype unspecified, MPD* = myeloproliferative disease specific subtype unspecified, NHL = non-Hodgkin lymphoma.

2.3.4 Family LK0139

LK0139 (shown in Figure 2.4) comprises a small family with two generations affected by HMs including an affected father-daughter pair, LK0139-001 and LK0139-005, both of whom have plasma cell disorders. A sibling (LK0139-003) of LK0139-005 was diagnosed with cervical cancer. Nieces of LK0139-001 are reported to have a HM and a brain tumour respectively. As shown in Table 2.5 and Figure 2.4 the father-daughter affected HM pair were genome sequenced and an unaffected daughter exome sequenced in this family.

Table 2.5 Characteristics of sequenced family members in LK0139.

Sample ID	HM Diagnosis	Sex	Age at diagnosis	Age at sample collection	Sample types available*	Sample type for NGS	Evidence of circulating disease in sample	WGS or WES
LK0139-001	PCT	M	57	81	PB	PB	No	WGS
LK0139-004	Unaffected	F	--	58	PB	PB		WES
LK0139-005	MM	F	52	53	PB / O	PB	No	WGS

* PB = genetic sample available from peripheral blood, O = genetic sample available from saliva collected in an Oragene kit.

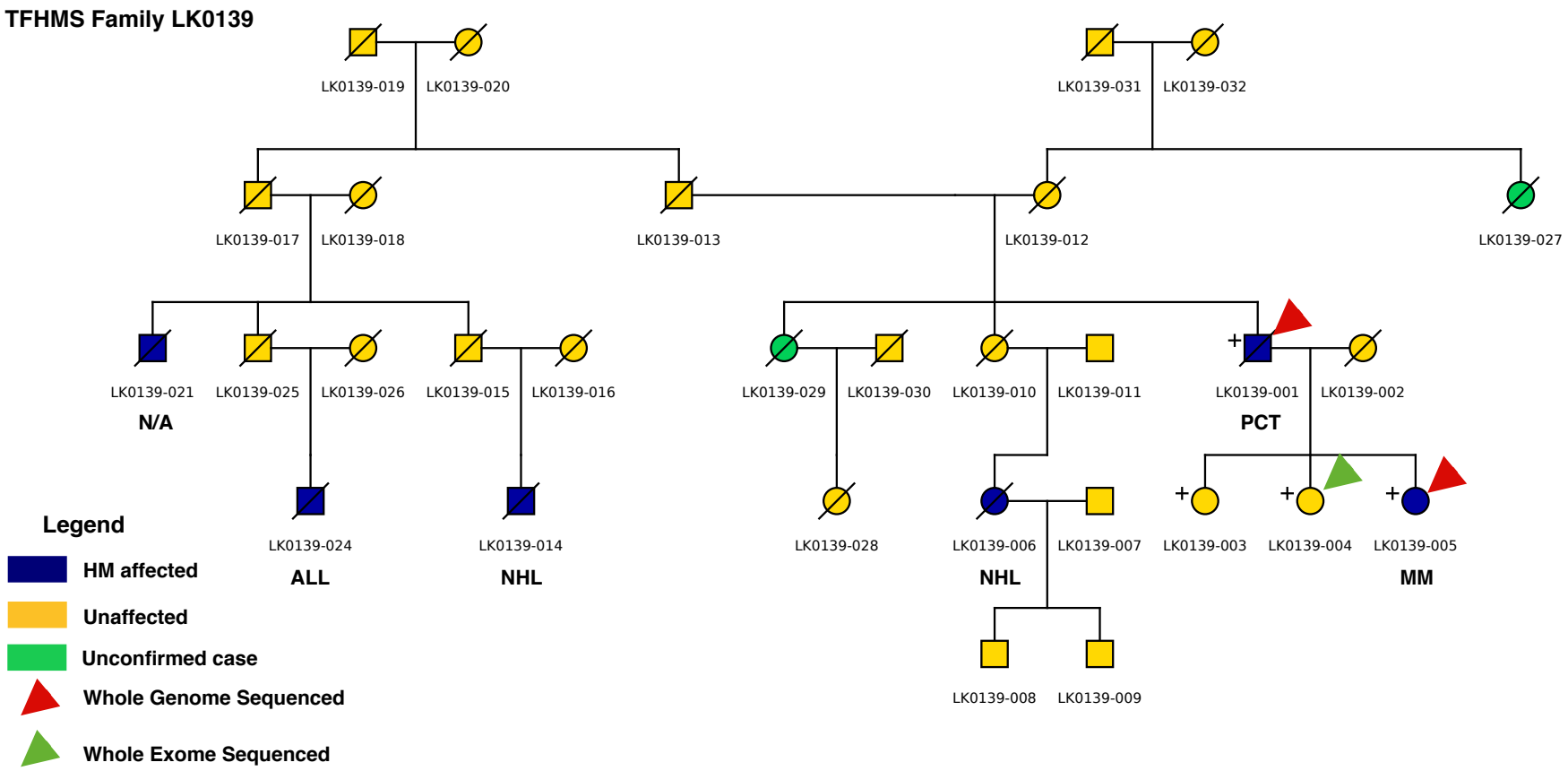


Figure 2.4 Extended pedigree of TFHMS family LK0139
For the current study both LK0139-001 and LK0139-005 were genome sequenced and LK0139-001's third unaffected daughter, LK0139-004, was exome sequenced as an informative control. Diagnosis abbreviations: ALL = acute lymphoblastic leukaemia, MM = multiple myeloma, NHL = non-Hodgkin lymphoma, N/A = diagnosis unavailable, PCT = plasmacytoma.

2.3.5 Family LK0153

LK0153 (shown in Figure 2.5) is a large TFHMS family that was ascertained from the original 1970s study[209-211]. There are ten HM cases across two generations with several different HM subtypes occurring including siblings, LK0153-003 and LK0153-004, who had mucosa-associated lymphoid tissue lymphoma (MALT lymphoma) and Waldenström macroglobulinemia (WM) respectively. As shown in Table 2.6 and Figure 2.5 the affected sibling pair and their father were genome sequenced. A separate but informative branch of the family without evidence of HM occurrence, were genome and exome sequenced as per Table 2.6 and Figure 2.5.

Table 2.6 Characteristics of sequenced family members in LK0153.

Sample ID	HM Diagnosis	Sex	Age at diagnosis	Age at sample collection	Sample types available*	Sample type for NGS	Evidence of circulating disease in sample	WGS or WES
LK0153-003	MALT	F	58	63	PB	PB	No	WGS
LK0153-004	WM / MDS / AML	M	55	61	PB / O	PB	Yes	WGS
LK0153-029	Unaffected	M	--	85	PB	PB	--	WGS
LK0153-086	Unaffected	M	--	70	PB	PB	--	WGS
LK0153-078	Unaffected	M	--	78	PB	PB	--	WGS
LK0153-079	Unaffected	F	--	73	PB	PB	--	WES
LK0153-084	Unaffected	M	--	46	PB	PB	--	WES
LK0153-080	Unaffected	M	--	49	PB	PB	--	WGS

* PB = genetic sample available from peripheral blood, O = genetic sample available from saliva collected in an Oragene kit.

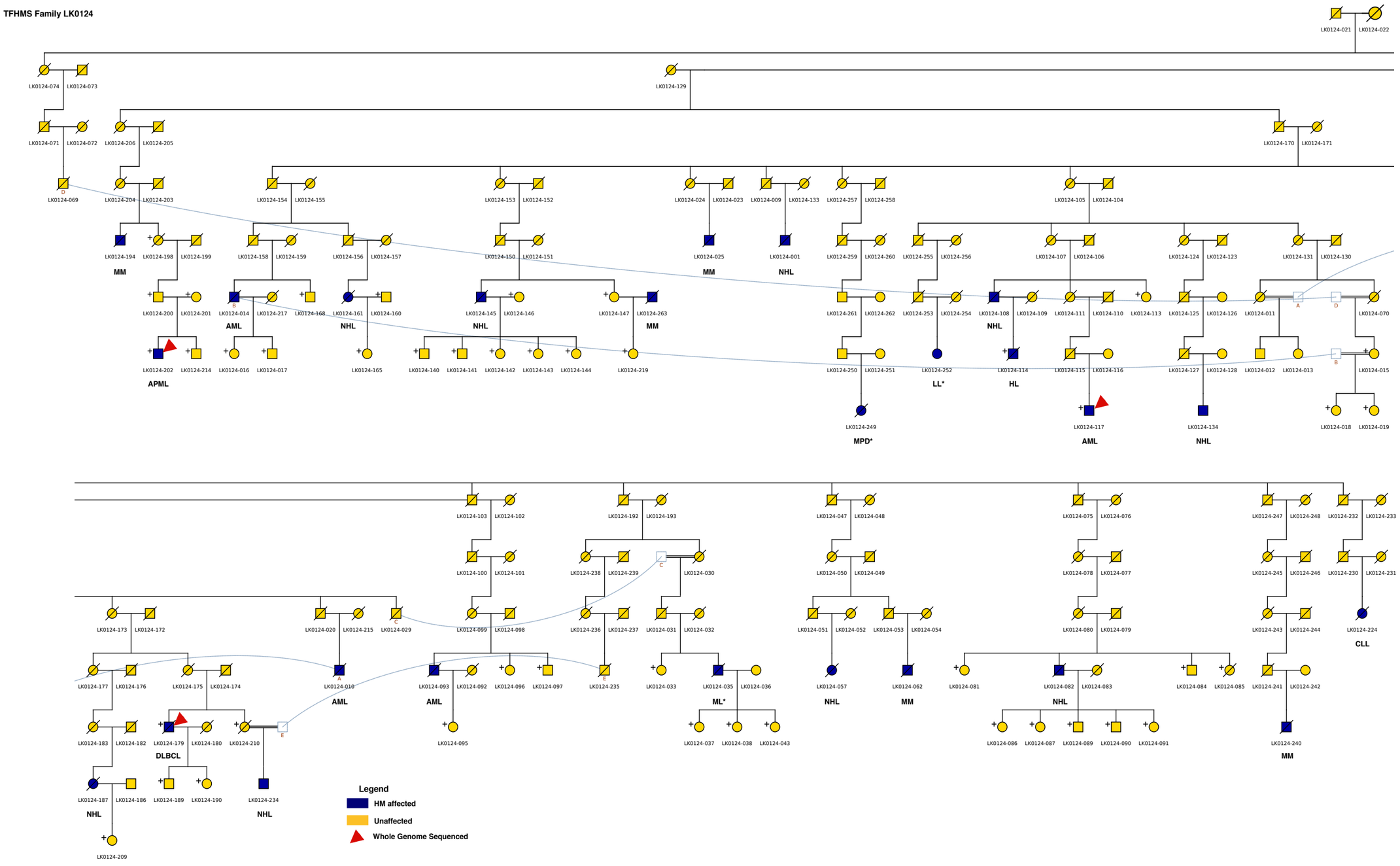


Figure 2.3 Extended pedigree of TFHMS family LK0124.
The focus in this family for the current study has been on three HM cases, LK0124-117, LK0124-202 and LK0124-179. These three cases are distant cousins. All were genome sequenced. Diagnosis abbreviations: APML = acute promyelocytic leukaemia, AML = acute myeloid leukaemia, CLL = chronic lymphocytic leukaemia, DLBCL = diffuse large B-cell lymphoma, HL = Hodgkin lymphoma, MM = multiple myeloma, ML* = myeloid leukaemia specific subtype unspecified, MPD* = myeloproliferative disease specific subtype unspecified, NHL = non-Hodgkin lymphoma.

2.3.6 Family LK2042

LK2042 (shown in Figure 2.6) is one of the largest HM families in the TFHMS. Members of this family were not ascertained from the original 1970s study. This family has five known generations affected by HMs totalling 33 cases with several different subtypes dispersed across the family. Additionally this family contains three consanguineous loops as a result of three cousin-cousin marriages, two of which were between first cousin pairs in the late 1800s and early 1900s. As shown in Table 2.7 and Figure 2.6 across this family six HM case were genome sequenced and two were exome sequenced. One unaffected relative was genome sequenced and three were exome sequenced.

Table 2.7 Characteristics of sequenced family members in LK2042.

Sample ID	HM Diagnosis	Sex	Age at diagnosis	Age at sample collection	Sample types available*	Sample type for NGS	Evidence of circulating disease in sample	WGS or WES
LK2042-003	T-cell ALL	M	7	26	PB / O	PB	No	WGS
LK2042-005	B-cell ALL	M	2	26	O	O	No	WES
LK2042-006	HL	M	29	37	PB / O	PB	No	WGS
LK2042-018	Unaffected	F	--	67	PB	PB	--	WGS
LK2042-231	DLBCL	M	58	66	PB	PB	No	WGS
LK2042-232	Unaffected	F	--	63	PB	PB	--	WES
LK2042-257	HL	M	66	67	PB / O	PB	No	WGS
LK2042-258	Unaffected	M	--	69	PB	PB	--	WES
LK2042-259	Unaffected	M	--	63	PB	PB	--	WES
LK2042-281	MCL	M	73	75	PB / O	PB	No	WGS
LK2042-290	ET	F	52	59	PB / O	PB	Unknown	WGS
LK2042-300	HL	M	61	64	PB / O	PB	No	WES

* PB = genetic sample available from peripheral blood, O = genetic sample available from saliva collected in an Oragene kit.

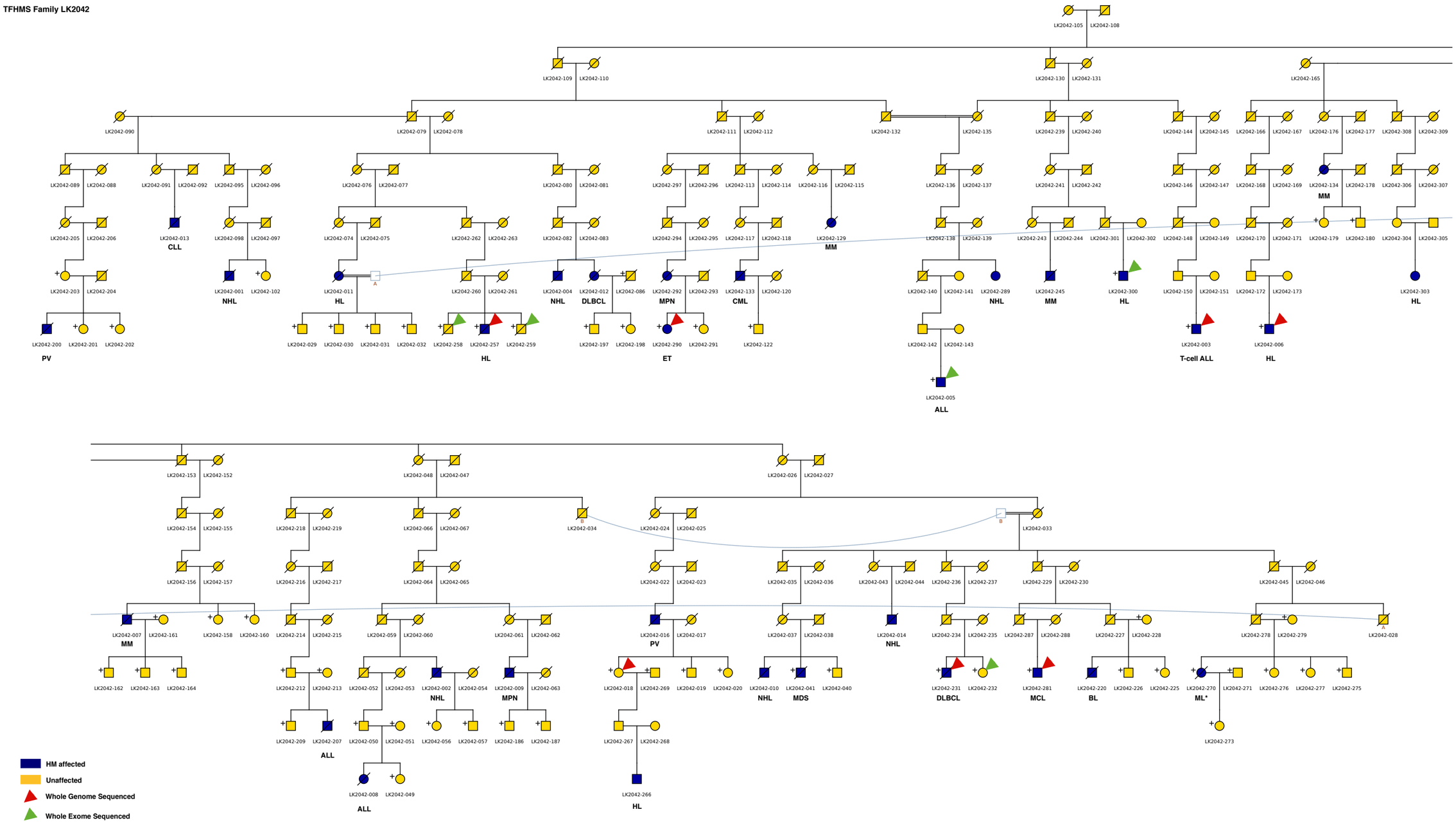


Figure 2.6 Extended pedigree of TFHMS family LK2042.
The focus in this family for the current study has been on eight HM cases, seven of which are still living, and four unaffected relatives who were selected as informative controls. The HM cases that were genome sequenced are LK2042-003, LK2042-006, LK2042-231, LK2042-257, LK2042-281 and LK2042-290. Two HM cases were exome sequenced, LK2042-005 and LK2042-300. Two unaffected siblings of LK2042-257, LK2042-258 and LK2042-259, and an unaffected sibling of LK2042-231, LK2042-232, were exome sequenced as informative controls. LK2042-018 who is the daughter of a HM case and the grandmother of another HM case was genome sequenced as a suspected obligate carrier. Diagnosis abbreviations: ALL = acute lymphoblastic leukaemia, BL = Burkitt lymphoma, CLL = chronic lymphocytic leukaemia, CML = chronic myeloid leukaemia, DLBCL = diffuse large B-cell lymphoma, ET = essential thrombocythaemia, HL = Hodgkin lymphoma, MCL = mantle cell lymphoma, MDS = myelodysplastic syndrome, MM = multiple myeloma, MPN = myeloproliferative neoplasm, NHL = non-Hodgkin lymphoma, ML* = myeloid leukaemia specific subtype unspecified, PV = polycythaemia vera.

2.4 Discussion

A comprehensive clinical review was conducted for all TFHMS HM cases. As HMs can present as a circulating tumour or a solid tumour-based malignancy, depending upon the subtype, clinical review permitted the identification of those with circulating tumour cells. It is likely that genomes extracted from samples largely comprising malignant cells will contain somatically acquired variations that would impact on the ability to identify true germline variants. However, the possibility that there are a significant population of circulating HM disease cells within collected samples is less important in light of this study's approach of identifying shared predisposing genetic variants. Finding shared genetic variants between multiple related HM cases provides support that these variants are germline and not somatically acquired.

Nevertheless, for sixteen of the eighteen sequenced cases it is unlikely that the collected blood samples contained a significant population of malignant cells. For the remaining two individuals the sample collected from LK2042-290 does contain circulating disease cells as a feature of her indolent HM subtype (essential thrombocythaemia) and for LK0153-004 it cannot be confirmed to the same level as other individuals whether malignant cells are present or not in the sample obtained.

This chapter also detailed the pedigree structures of the study families. The sequenced cases and unaffected relatives from these five TFHMS families were selected to permit a number of different analysis strategies aimed at identifying shared variants predisposing to HMs. This is inline with the study hypothesis, as discussed in Chapter 1, that familial cases will be enriched for shared variants predisposing them to HMs. The caveat of predisposition is that unaffected relatives may also carry these variants without developing disease. Reduced penetrance of complex diseases such as HMs must be taken into consideration, as not all mutation carriers will develop disease[216].

It is important to note that the TFHMS families studied here are multigenerational Tasmanian families that can be genealogically traced back to the approximately 10000 Tasmanian founder families as discussed in section 2.1.1. As has been seen in other known founder populations, such as Finland, there can be a high frequency of rare and/or novel variation present within the general population, due to the

bottleneck effect several generations earlier when the population began from a smaller number of individuals or due to limited regional population movements[217]. When the frequency of genetic variants identified in these populations is compared to broader population resources such as the 1000 genomes project, the presence of variants that occur at low frequencies in broader resource (i.e. are rare) are seen at higher frequencies in the founder population (i.e. are common). The more frequent occurrence of such rare variants in founder population necessitates an analysis design that utilises founder population matched controls. The historically documented population bottle necks in Tasmania are not as pronounced as other isolated populations such as Finland[217], so while the 'common rare variant' issue is an important consideration it should not be a major confounder in this study.

In comparison to other published familial HM studies the TFHMS has similarities to but also important differences to other resources. Recent publications that have used NGS in HM families have concentrated on using families with single related subtypes of HMs to identify susceptibility variants, for example *GATA2* variants in familial MDS/AML[148] and *PAX5* variants in familial ALL[189]. Instead, the TFHMS families have multiple subtypes of disease, such as in family LK0051 where a T-cell and two differing B-cell HMs were genome sequenced, or in the LK0153 family where two different B-cell malignancy subtypes were genome sequenced. These families provide opportunities to question whether there is a common genetic predisposition to multiple subtypes. Similar to other familial HM studies, TFHMS families include examples of first-degree relative case pairs, such as in families LK0153 and LK0139, however the extensive genealogical research conducted using this resource and the benefit of using an isolated population has allowed identification of larger HM affected families, such as in families LK0124 and LK2042. NGS of large HM families with cases more genealogically distant than first or second cousins have not been reported, whereas here three of the five families studied have more distant affected relatives. This provides a novel opportunity to identify predisposing genetic variants shared between distant relatives.

Chapter 3 - Genome and exome sequencing, variant identification and prioritisation

3.1 Introduction

Since the development of NGS, an area of ongoing change has been the bioinformatics techniques used for sequence alignment to the reference genome and subsequent identification and analysis of sequence variation. A recent report comparing five different alignment and variant calling pipelines showed low concordance between pipelines for single nucleotide variant (SNV) and insertion / deletion (indel) variant identification[218]. It is apparent from this study that the use of an established bioinformatics pipeline for alignment and variant identification is crucial. With several comparable options available, this dissertation has used, primarily for speed and computing RAM requirements[219], the Burrows-Wheeler alignment (BWA) method[220], with additional post alignment processing using the Genome Analysis Tool Kit (GATK)[221]. SAMtools[222] was chosen for SNV and indel identification because of the ability to identify variants across multiple samples simultaneously.

A strength of this project lies in the family-based study design providing the ability to construct sharing-based analyses within families to identify variants segregating with disease. However studying a disease by NGS in related individuals will identify many shared variants by virtue of kinship that are not related to the disease. Accordingly a methodology to identify the variants that are most likely to contribute to disease susceptibility in affected individuals must be established. As it is not yet feasible to examine or validate the occurrence of hundreds of likely susceptibility variants, a prioritisation strategy must be used to reduce and focus analyses on a specific variant set. In this study variant analysis to prioritise those likely to be contributing to disease was conducted using two approaches. One approach applied was a probabilistic approach using statistical prediction algorithms to prioritise the most likely candidate variants contributing to disease. The primary approach used however, was a heuristic approach using manual filtering of annotated variants with a series of targeted assumptions to narrow down variants to a small number for further follow-up.

An example of a probabilistic tool is pVAAST[223] (developed by the Huff Lab at the MD Anderson Cancer Center), a variant prioritisation tool that builds upon VAAST[224,225] (from the Yandell Lab at the University of Utah), the probabilistic disease gene finder tool. By incorporating pedigree information pVAAST conducts a linkage-based analysis, together with VAAST's rare variant association test, searching for rare, functional variants inherited in the family that could be contributing to disease. pVAAST was chosen because it is able to identify disease causing variants in family-based studies of conditions that have recessive or dominant inheritance patterns as well as being able to handle challenges such as incomplete penetrance and locus heterogeneity[223]. Such an unbiased approach is very favourable when analysing a complex disease. Here the use of pVAAST was explored in an analysis of the LK0051 family.

One widely used tool for facilitating heuristic-based variant filtering is ANNOVAR[226], originally designed by Dr Kai Wang. A heuristic-based strategy for variant prioritisation requires development of specific analysis hypotheses. An example would be to identify variants shared between all cases in the family but not present in related controls. Such hypotheses then rely on the annotation of variants from the wealth of publically available data such as population frequencies from large NGS consortia projects, predictions of variant effect and locations of the variant within potentially functional genomic regions. Then, based on the specific hypothesis the annotated variants are filtered and reduced to a smaller set of variants. ANNOVAR facilitates this annotation and filtering, forming the foundations of the tiered prioritisation strategy that was used in this project to identify variants likely to contribute to disease in the five families sequenced.

Once a filtered set of variants, conforming to the specific hypothesis, has been identified the analyses can move from a variant-based to a gene-based approach. Gene-based analyses are directed at prioritising variants based upon what is known about the gene in which the identified deleterious variant occurs. This information can include gene expression profiles in normal tissues or cancer cells, known literature interactions with other molecules, as well as the known function of the gene and its potential to have a biological relationship with HM disease when that function is

altered. The gene-based analyses presented here have relied specifically on two complementary tools: the publically accessible Phevor[227] (from the Yandell Lab at the University of Utah), which is a phenotype driven ontology-based gene ranking tool; and a commercial tool, QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity), which is a tool that builds interaction networks between molecules based on a curated literature database of interactions.

Specifically, Phevor uses a set of defined phenotype terms related to the disease of interest to combine outputs from multiple biomedical ontologies including the Human Phenotype Ontology, the Mammalian Phenotype Ontology, the Disease Ontology and the Gene Ontology[227]. Phevor connects these outputs together, creating a framework built around the defined phenotype. The queried gene list is then reprioritised in light of this knowledge, the gene's connections into the ontologies and the corresponding phenotype defined. Complementary to this is QIAGEN's IPA®, which is a commercial resource of curated literature-based molecular interactions. Using IPA® the queried gene list can be searched for direct molecule-to-molecule interactions within and between other lists of known HM related genes. Using Phevor and IPA® to connect genes with deleterious variants identified in the prioritisation analysis to genes known to be involved in HM development, or to the HM phenotype, increases the evidence implicating the identified variants in disease development.

The final stage of a heuristic-based analysis of variants is to assess the variants more closely to predict their direct effects on gene function, at the transcript or protein level, to identify those most likely to contribute to HM development.

3.2 Aims

The objective of these analyses was to establish a method for identifying susceptibility variants in the five TFHMS families examined employing either a probabilistic or heuristic approach.

A probabilistic strategy, using pVAAST[223], was trialled using the sequence data in family LK0051. However, this was found to have a number of limitations in application to this data, so instead, a heuristic-based filtering approach with a tiered

family-by-family strategy was used. This approach, initially based at the variant level, focuses on variants that are likely to be deleterious and in gene regions where *in silico* tools can be used to predict the variant effect. Then the approach moves to the gene level. At the gene level prioritisation is based on the known biology of the gene, its expression patterns, protein functions and interactions. Variants in genes that have a biological indication that if functionally disrupted could contribute to HM development are prioritised for further analysis. An additional analysis for each family was to examine non-shared variants specific to each HM case.

3.3 Methods

3.3.1 Generating the WGS and WES data

Due to the availability of resources and the schedule of grant funding, WGS and WES were performed over a four-year period. All samples were sequenced on the Illumina platform.

Prior to sequencing, the selected DNA samples were quantitated using the Qubit 2.0 fluorometer system with both broad range and high sensitivity kits where required, according to the manufacturer's protocols. Purity was determined by measurement on the NanoDrop 1000 using the $A_{260/280}$ ratio. Samples with purity outside a 1.7 - 1.9 range were processed by phenol chloroform extraction and ethanol precipitation (Appendix 3.1) prior to NGS analysis. DNA for NGS analysis was also visually examined by agarose gel electrophoresis to assess for sample degradation.

For WGS 3 - 5 µg of DNA (100µL at 30 - 50 ng/µL) with $A_{260/280}$ ratio in range 1.7 - 1.9 was used for 100 bp paired end genome sequencing to a mean coverage of 30× or greater. For WES 30 ng of DNA (20 µL at 1.5 ng/µL) with $A_{260/280}$ ratio in range 1.7 - 1.9 was sequenced in-house using the Nextera Exome Enrichment System and the Illumina TruSeq paired-end cluster kit v2 for 100 bp paired end exome sequencing to a mean coverage of 30× or greater.

3.3.2 Genome and exome sequencing alignment

An analysis pipeline, outlined in Figure 3.1, was established for consistent alignment and variant calling of the genomes and exomes on the 8000 processor compute cluster MEDUSA housed at Texas Biomedical Research Institute, San Antonio, Texas, USA with the assistance of their lead NGS bioinformatician Mr Juan Peralta.

WGS data was obtained from Illumina in BAM file format. Each BAM file was converted to FASTQ format to obtain de-multiplexed paired end sequencing reads of 100 bp in length. For WES data, bases were called from HiSeq 2500 raw intensity files using Illumina's CASAVA 1.8+ suite[228] and emitted in FASTQ format of 100 bp paired-end sequencing reads.

A custom mapping pipeline built with GNU Make[229] was applied to both genomes and exomes, as shown in Figure 3.1. Use of GNU Make facilitates the scaling of analysis pipelines from one to multiple samples, using one to thousands of processors on the MEDUSA compute cluster. To facilitate parallelisation across the cluster FASTQ files were split into fragments containing 4,000,000 sequencing reads. Reads were aligned to the hg19 (GRCh37, February 2009) version of the human reference genome using BWA (v0.6.1)[220]. BWA was chosen over other aligners primarily for its speed. Fragmenting the alignment into smaller tasks results in many tasks with fewer reads to align, leading to shorter overall run times instead of a few tasks with many reads to align and longer run times. Aligned reads were then processed with a custom awk script (by Mr. Juan Peralta, Appendix 3.2) to ensure consistency of the mapped metadata and compliance with the SAM/BAM standard. The resulting alignments were converted to BAM format using SAMtools[222]. BAM fragments were merged and pooled on a per sample basis and SAMtools (v0.1.18)[222]was used to mark likely PCR duplicates. The IndelRealigner method from GATK (v1.6)[221] was used for realignment around indels and GATK's BaseRecalibration method was used for base quality score recalibration. This generated realigned, recalibrated and sorted BAM files for each genome and exome.

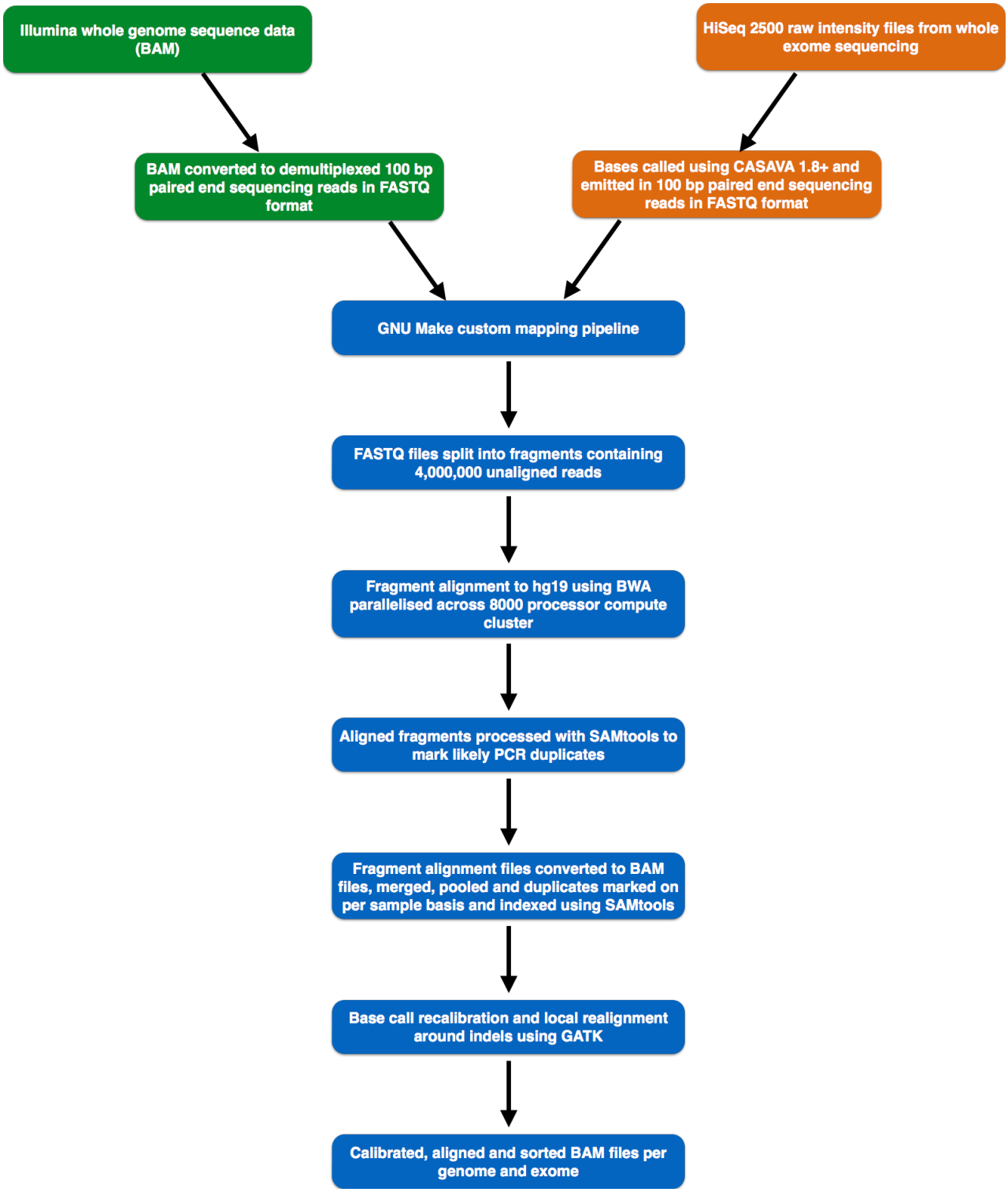


Figure 3.1 Pipeline schematic for WGS and WES alignment to hg19 reference genome.

3.3.3 WGS and WES quality assessment

For quality assessment the final aligned genome BAM files were assessed using FASTQC (v0.10.0)[230] and additionally for exomes, Qualimap (v2.0)[231]. Mean WGS coverage was calculated using SAMtools idxstats (v0.1.18)[222] and mean WES coverage depth was calculated using Qualimap (v2.0)[231]. Plots of genome and exome coverage were calculated using BEDTools genomeCoverageBed (v2.17.0)[232] and plotted using R (v3.1.0)[233].

Exomes were assessed in all instances using the relevant TruSeq Exome Enrichment Kit targeted regions from Illumina from:
http://support.illumina.com/downloads/truseq_exome_targeted_regions_bed_file.html

3.3.4 Single nucleotide and indel variant calling

SAMtools mpileup (v0.1.18)[222] was used for variant calling of single nucleotide variants and indels. To parallelise variant calling across the compute cluster the genome was fragmented into 1,125 × 3,000,000 base pair regions. Using BEDTools subtractBed[232] the 1,125 regions were intersected with 457 regions from UCSC's gap track[234] and a set of 411 blacklist regions from the Broad Institute[235] to remove genomic regions with low mappability, unusually high mappability, or telomeric and centromeric regions. This resulted in 1,737 genomic regions to call variants across using SAMtools mpileup. The resulting VCF files generated were further processed by GATK (v1.4.37) to add quality filter tags to each sample's variants to facilitate filtering of variants on quality at a later analysis stage. Variants with coverage depth ≥ 10 and quality scores ≥ 20 were called high quality variants. Variants meeting one but not the other quality criterion were called medium quality variants. Variants failing to meet either quality criteria were called low quality variants. Any variant that was tagged as a high quality variant in at least one sample was carried forward into the final VCF files. These variants were used in the probabilistic pVAAST analyses described below. Example SAMtools mpileup and GATK commands for one region are shown in Appendix 3.3.

3.3.5 ANNOVAR SNV and indel filtering and annotation

SNV and indel variants in VCF file format were converted into ANNOVAR[226] input format using a custom script (Appendix 3.4). Using the variants_reduction.pl script of ANNOVAR and the allele frequencies from the 1000 Genomes Project (1000GP) Caucasian population (Phase 1 v3 April 2012 release)[236], common variants with an alternate allele frequency in 1000GP of >1% were removed from further analysis. Variants were then annotated with human genome RefSeq[237] designations (as at 12/07/2014) which provided gene and gene region-based annotations, Table 3.1 describes the designation priority order as used by ANNOVAR. Annotated variants located within intronic or intergenic regions were removed from further analysis leaving variants present in exonic, splicing, 3' and 5' untranslated regions (UTR), upstream, downstream and within non-coding RNA (ncRNA).

Table 3.1 ANNOVAR RefSeq genomic region annotation definitions[226].

Annotation	Definition	Precedence
Exonic	Variant is within gene coding exon	1
Splicing	Variant is within 2-bp of exon / intron boundary	1
ncRNA	Variant is within a known transcript that is non-coding	2
5'UTR	Variant is within 5'UTR gene region	3
3'UTR	Variant is within 3'UTR gene region	3
Intronic	Variant is within a gene intron	4
Upstream	Variant is within a 1000 base pair region upstream of transcription start site	5
Downstream	Variant is within a 1000 base pair region downstream of transcription end site	5
Intergenic	Variant is in an intergenic region	6

These variants (exonic, splicing, 3' and 5' UTR, upstream, downstream and within ncRNA) were then annotated with information contained within the databases listed in Table 3.2. Most were built and sourced from ANNOVAR, where the table_annovar.pl script was used, but a number are custom annotations or accessed through specific web-based tools as described in Table 3.2. Specifically the Caucasian data from the most recent 1000 Genomes Project data release (September 2014, Phase 3 v5), together with allele frequency data from the UK10K consortia[238] was applied to the variants to again remove those with alternate allele frequencies ≥ 1% in these large scale resources. This resulted in variants with alternate allele frequencies < 1%

for the next stage of analysis. These variants were used in family specific analyses using the heuristic filtering approach to prioritise variants for follow-up using a tiered strategy.

Table 3.2 Databases accessed for variant-based annotations.

Database Name	Database Source	When variant is identified in database annotates variant:	Database Reference
RefGene	ANNOVAR	With RefSeq gene ID, genomic region, exonic function (synonymous, nonsynonymous, frameshift, stop loss, stop gain), amino acid changes	237
UCSC CpG Islands	UCSC	When located in a CpG island	234
UCSC cytoBand	UCSC / ANNOVAR	With macroscopic chromosome metaphase cytoband location	234
TargetScanS	UCSC / ANNOVAR	With TargetScanHuman 5.1 generated prediction of whether variant is located within a miRNA binding site in 3'UTR of genes	239
encode_tfbs	UCSC / ANNOVAR	With location of transcription factor binding sites that are determined from the ENCODE project ChIP-Seq experiments	240
COSMIC 70	COSMIC / ANNOVAR	With information from COSMIC when variant is previously observed in cancer genomes, what tissue and frequency	109,241
ljb23_all	dbNSFP / ANNOVAR	Annotates variants with a variety of functional prediction and conservation scores from the following tools: whole-exome SIFT scores, PolyPhen2 HDIV scores, PolyPhen2 HVAR scores, LRT scores, MutationTaster scores, MutationAssessor score, FATHMM scores, MetaSVM scores, MetaLR scores, GERP++ scores, PhyloP scores and SiPhy scores.	242
dbSNP 138	NCBI / ANNOVAR	With dbSNP RS identifiers	243
CADD	http://cadd.gs.washington.edu	With CADD scores and Phred-scale like conversion of CADD scores, a deleteriousness score for SNVs and indels	244
SilVA	http://compbio.cs.toronto.edu/silva	With a functional prediction score and a prediction label (for exonic synonymous SNVs only)	245
ucscGenePfam	UCSC	With manually curated protein domain information from the Pfam (Protein family) database (for exonic SNVs and exonic indels)	246
1000 Genomes Project Caucasian allele frequencies (Phase 3v5 September 2014 release)	1000GP / ANNOVAR	With alternate allele frequencies for 507 Caucasian samples in 1000GP cohort	236
UK10K consortia Caucasian allele frequencies from whole genome cohorts	UK10K	With alternate allele frequencies from 4000 whole genomes in the UK10K ALSPAC and TWINSUK cohorts	238

3.3.6 A probabilistic approach to disease gene identification using pVAAST

The pVAAST[223] software is the pedigree-based adaptation of the VAAST[224,225] software, a probabilistic disease gene finder. For each family pVAAST takes as input a pedigree structure and disease information about family members, a target set of variants from the family being analysed, a background set of variants from population and bioinformatics methods matched controls, a configuration file detailing specific settings for the analysis (Appendix 3.5) and a file defining gene features; refGene_hg19_with_introns.gff3 sourced from the VAAST server was used.

To trial pVAAST the LK0051 family (target) was analysed together with three background control sets. The first was the default VAAST background sourced from the VAAST server. This is a background generated from all genome data from all available populations in the 1000GP dataset as at December 2011. The specific file used was: 1KG_refGene_Dec2011_CGDiv_NHLBI_NoCall.cdr. The second was a background constructed from 100 Caucasian samples from the 1000GP. For this background 70 Caucasian exomes and 30 Caucasian genomes from the 1000GP data set were downloaded and variants were multi-sample called as per 3.3.4. The third background was constructed from the remainder of the TFHMS samples used in this study (N=26, excluding the LK0051 samples as they were used as the target).

Then pVAAST was run using the following command settings:

```
VAAST –m pvaast –e ––indel ––rate 0.01 –o output_name –pv_control
ConfigurationFile.ctl –p 55 –gw 1e6 refGene_hg19_with_introns.gff3
BackgroundFile.cdr
```

This command calls the VAAST program and specifies that the method (`–m`) to use is the pvaast analysis, to score all variants both coding and non-coding and indels (`–e ––indel`), with the expected maximum disease allele frequency observed in the background set at 1% (`––rate`), to use the specific configuration file (`–pv_control`, as per Appendix 3.5) and to parallelise the run across 55 threads (`–p`) and to perform 1×10^{6} analysis permutations (`–gw`), meaning the minimum P-value that can be achieved is 1×10^{-6}.

Multiple pVAAST analyses were completed using a full set of variants in the LK0051 target and also an analysis with variants restricted to a minor allele frequency ≤1%, filtered based on the 1000GP and UK10K Caucasian population frequencies, excluding intergenic and intronic variants.

The VAAST PDF reporter script (`vaast_pdf_reporter.pl`) was used to generate Manhattan and quantile-quantile plots.

3.3.7 A tiered heuristic approach to disease variant and disease gene identification

A tiered heuristic approach, following family-based analysis strategies, was developed and used to identify and prioritise candidate variants and genes that may have predisposed HM affected family members to disease development.

3.3.7.1 Tier One: Family specific variants

In the first step of this approach the initial filtered variants identified across the 31 sequenced individuals were reduced to an analysis set by removing high quality variants (with coverage depth ≥ 10 and quality scores ≥ 20) variants present in relatives of HM cases who were unaffected and non-obligate carriers, across the five families. A limitation of this is that lower quality true variants in these control individuals may be retained in the analysis set, however removing variants that occur at lower qualities in control samples from the cases risks removing true causal variants. Then a tiered strategy was developed and applied to each family individually. Not all families carry each of the analysis set variants so a smaller number of variants are applicable in each family.

3.3.7.2 Tier Two: Family specific sharing

After removal of variants present in unaffected non-obligate carrier relatives, each family has different analysis opportunities. Sharing analyses were designed specifically for each family but were all based on identifying shared variants between related cases.

3.3.7.3 Tier Three: Deleterious variants according to the CADD model

To score variants based upon their predicted deleteriousness, i.e. the potential for a variant to have an effect on gene function, Combined Annotation-Dependent Depletion (CADD) phred-like scaled C scores[244] (v1.1) were applied to identify variants likely to be damaging. CADD measures deleteriousness of single nucleotide and small insertion / deletion variants by combining in a single score 63 variant annotations related to variant function, including nucleotide and amino-acid conservation across species and epigenetic regulatory information from ENCODE[244]. A CADD phred-like scaled C score threshold of ≥ 10 was used to select for variants predicted to be deleterious. These are variants with the highest 10% of all scores, the top 10% predicted to be most deleterious of all possible variants under the CADD model. For variants in 5'UTRs or 3'UTRs a scaled C score of ≥ 20 was used to identify the most deleterious regulatory variants.

3.3.7.4 Tier Four: Deleterious variants with the most interpretable impact upon gene function

Interpreting the potential effects of a range of variants on the function of a gene is challenging. A variant can be predicted to be highly deleterious based on CADD scoring, but have no foreseeable effect on gene function. To this end, to prioritise variants that are able to have a biological prediction as to their effect on gene function variants were filtered to:

- Nonsynonymous exonic variants located in known protein family domains, according to the Pfam database[246]
- Synonymous exonic variants predicted to be deleterious by an additional tool, SilVA[245]
- Variants located in splicing regions
- Upstream and 5'UTR regulatory variants in a known transcription factor binding site based on ChIP-Seq data from ENCODE[247], specifically the 'Txn Fac ChIP V2' data from the UCSC genome browser[234] with CADD scaled C scores ≥ 20
- Downstream or 3'UTR regulatory variants in predicted miRNA binding sites based on the TargetScanHuman 5.1 database[239]

3.3.7.5 Tier Five: Network-based prioritisation of genes

To draw a stronger connection between the prioritised variants and HMs Tier Five moved from a variant-based analysis to a gene-based analysis to prioritise genes likely to be involved in HMs. Phevor[227] and IPA® were used.

In Phevor a HM related phenotype was defined using the terms: haematological neoplasm, hematopoietic or lymphoid organ development, hematopoietic stem cell differentiation, hematopoietic progenitor cell differentiation, hematopoietic stem cell proliferation, hematopoietic stem cell homeostasis, cancer, hematologic cancer, and hematopoietic system disease. Genes identified in Tier Four were queried against this phenotype through Phevor, with default filtering conditions enabled. Genes not filtered by Phevor were classified as having some evidence of relationship to the HM phenotype defined and were prioritised for further analysis.

For gene prioritisation using IPA®, a background of HM related genes was manually curated from the literature. This included genes recurrently somatically mutated in HMs (as per Chapter 1 Table 1.8[171]) and genes identified as contributing to a germline predisposition to HMs (Chapter 1 Table 1.5, Table 1.6 Table 1.7[111] and Table 1.9). This list of HM related genes was entered into IPA® to create a HM gene background network, as shown in Figure 3.2. Each of these interactions, shown with edges in Figure 5.2, is a direct molecule-to-molecule interaction that has been curated from the literature and is contained within the Ingenuity® Knowledge Base. Each interaction, including its direction, has been curated from the biomedical literature or third-party databases by trained curators[248]. There are approximately five million connections within the Knowledge Base and together this builds a network of approximately 40,000 nodes (mammalian genes and their products) with approximately 1,480,000 edges between them, representing experimentally observed relationships[248]. The background network formed in this analysis was used to identify connections between the genes from each Tier Four prioritisation and this background network of genes known to contribute to HM development. Within IPA® the pathway 'Connect' tool was used, specifying 'direct' interactions only with 'experimentally observed' or 'high (predicted)' confidence levels, other settings were used at default. The variants in genes that were connected into the HM background network by this tool could potentially have a role in HM predisposition and development.

An additional network-based analysis was conducted using the Network of Cancer Genes 4.0 (NCG 4.0) database[249]. The NCG 4.0 analysis was used to highlight genes that have been previously identified as mutated in cancer and to identify genes known to be false positives in NGS projects, such as *TTN* and *OBSCN* two of the largest coding genes in the genome, which due to their long length are more likely to have mutations than other genes. Genes in each analysis that were prioritised by Phevor and/or IPA®, and not predicted to be possible false positives by NCG 4.0, were carried forward to the next Tier analysis for each family.

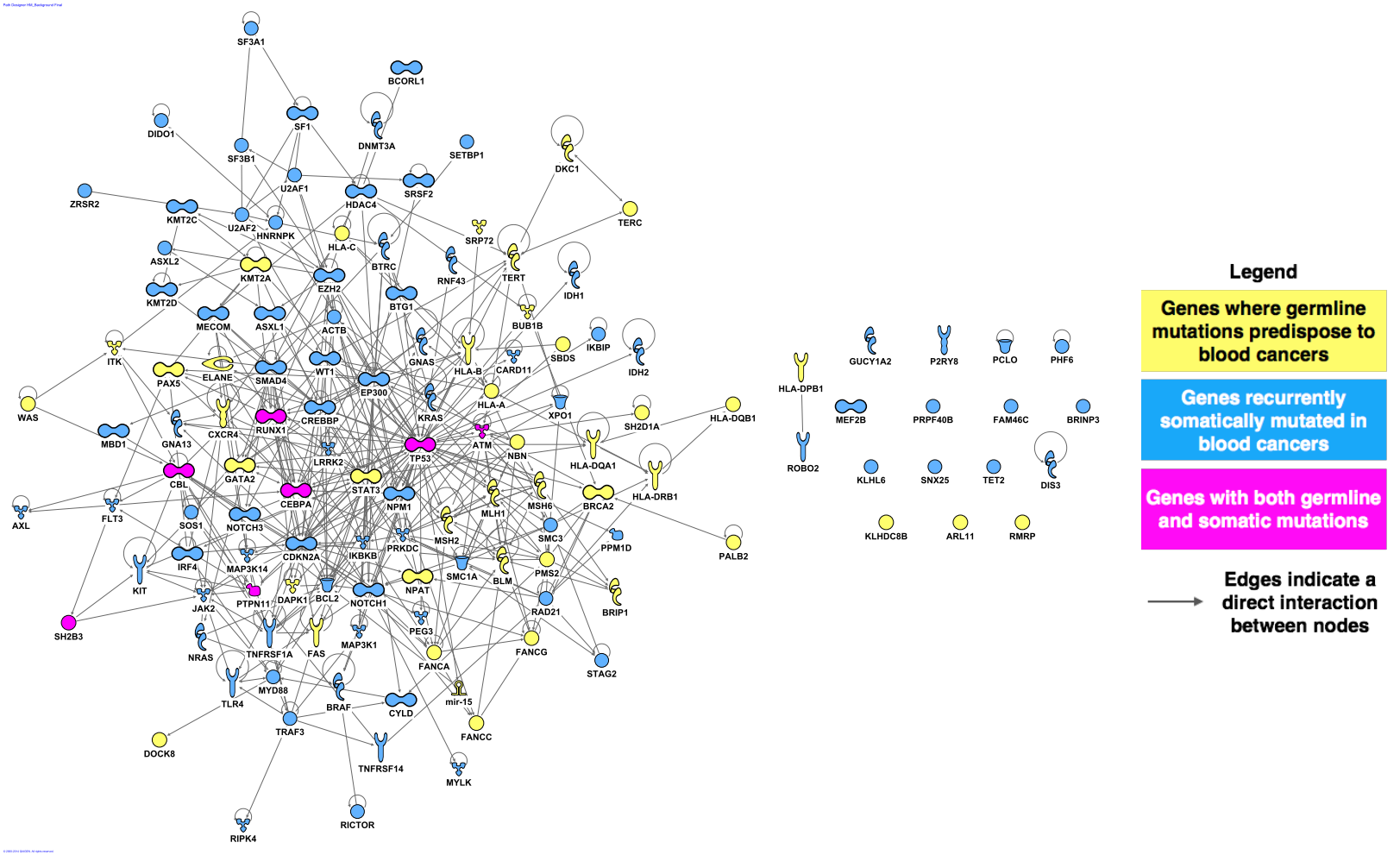


Figure 3.2 Ingenuity Pathway Analysis network of HM background genes curated from the literature
This network was formed from direct molecule-molecule interactions identified within the Ingenuity knowledge base, a curated database of interactions. Genes are shown as nodes and gene / gene product interactions as edge connections between nodes, as shown in the legend. Seventeen genes do not connect in with the larger network, as indicated to the right of the network diagram. Seven genes, indicated in magenta, have been shown to both contribute to an inherited predisposition to HMs and be recurrently mutated somatically in HMs.

3.3.7.6 Tier Six: Prioritisation based on analysis of known gene functions, expression and predicted effects of variant on protein function

For shared, network prioritised variants identified in each family, Tier Six examines the predicted effects of each variant in the context of gene function and gene expression profiles. In Tier Four nonsynonymous exonic variants in Pfam domains were selected. Many Pfam domains have a solved protein crystal structure or a structure that can be modelled based on homology to other known protein structures from humans or other species. This allows for the specific amino acid change to be examined in the context of the protein itself.

Project HOPE (Have yOur Protein Explained)[250] was used to model effects of single amino acid changes on protein function. Particularly, conservation of the amino acid was used, as changes in highly conserved amino acids are more likely to be deleterious to protein function. When the amino acid was not conserved, and the exact variant amino acid or an amino acid with similar properties occurred in homologous protein sequences in other species, this variant was not prioritised further. Project HOPE is limited to modelling single amino acid changes from SNVs. Insertion and deletion variants, causing coding frameshift amino acid mutations in Pfam domains, were evaluated manually by examining amino acid conservation and involvement of affected amino acids in protein function (e.g. trimer interface residues, disulphide bond residues).

Synonymous exonic variants that were predicted to be deleterious using SilVA were evaluated for potential effects on codon usage bias or proximity to splice sites. Variants in splicing regions were specifically assessed to examine their predicted effects on the key splice site nucleotides as described in Baralle and Baralle[251]. Regulatory variants upstream / 5'UTR in transcription factor binding sites based on ENCODE data and downstream / 3'UTR in miRNA binding sites based in TargetScanS predictions were carried forward without further assessment.

Gene expression profiles were then examined for each gene using publically available microarray data presented in BioGPS[252], in particular using the U133A gene atlas human gene expression data of 79 human tissues, including expression from 15

tissues and five cell lines relevant to HMs, that is described in Su *et al*.[253]. A finding of striking gene expression in haematopoietic or lymphoid tissues, or HM cell lines was additional evidence in gene prioritisation. Reported gene literature, and gene mutations recorded in COSMIC[108-110], were also assessed to evaluate the functional potential of the gene in HM development.

As the overarching hypothesis for this project is that shared variants contribute to disease in these individuals, non-shared variants were not prioritised at this stage for further follow-up.

3.4 Results

3.4.1 Genome and exome sequence quality assessment

All genomes and exomes passed quality and coverage assessment. Representative FASTQC reports for LK2042-003 (WGS) and LK2042-005 (WES) are located in Appendix 3.6. Mean WGS coverage results from analysis with SAMtools idxstats (v0.1.18)[222] and mean WES coverage depth from analysis with Qualimap (v2.0)[231] are shown in Table 3.3. A representative WES Qualimap report for LK2042-005 is located in Appendix 3.6. Plots of genome and exome coverage calculated using BEDTools genomeCoverageBed (v2.17.0)[232] are shown in Figure 3.3.

Table 3.3 Mean coverage depth results for genomes (using SAMtools) and exomes (using Qualimap).

Sample	NGS method	Mean coverage
LK0051-001	WGS	39.56
LK0051-007	WES	34.12
LK0051-128	WGS	35.79
LK0051-159	WGS	37.06
LK0051-165	WES	27.05
LK0124-117	WGS	54.20
LK0124-179	WGS	47.43
LK0124-202	WGS	48.70
LK0139-001	WGS	46.81
LK0139-004	WES	38.59
LK0139-005	WGS	47.61
LK0153-003	WGS	46.94
LK0153-004	WGS	42.68
LK0153-029	WGS	45.07
LK0153-078	WGS	42.24
LK0153-079	WES	46.72
LK0153-080	WGS	43.95
LK0153-084	WES	26.52
LK0153-086	WGS	41.07
LK2042-003	WGS	40.40
LK2042-005	WES	34.17
LK2042-006	WGS	38.83
LK2042-018	WGS	36.96
LK2042-231	WGS	39.42
LK2042-232	WES	36.58
LK2042-257	WGS	37.59
LK2042-258	WES	44.79
LK2042-259	WES	37.09
LK2042-281	WGS	35.64
LK2042-290	WGS	37.51
LK2042-300	WES	39.82

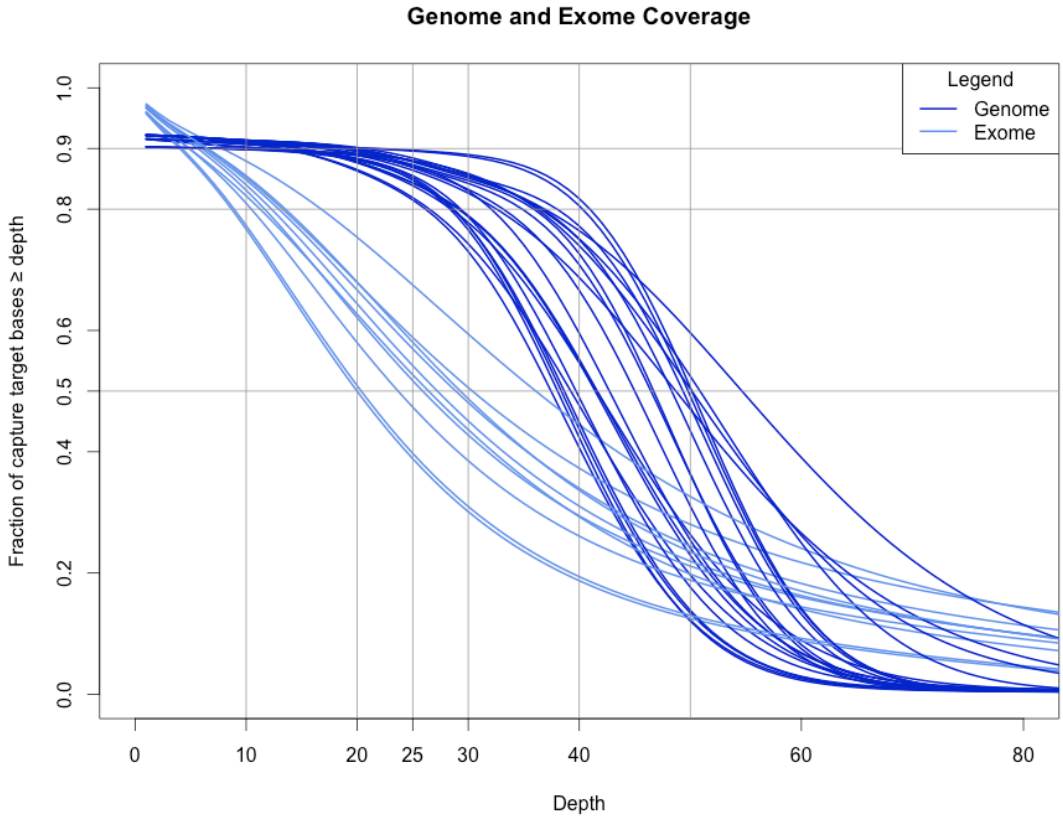


Figure 3.3 Genome and exome coverage plot.
Calculated in BEDTools and plotted in R this plot shows the fraction of bases (Y-axis) at a coverage depth (X-axis). Genomes have high coverage across genome, with ≥ 80% of genome with depth ≥ 20× read depth. Exomes have lower overall coverage of exome target regions with approximately 80% of exome coverage at ≥ 10× read depth.

3.4.2 Variant identification and initial filtering

The total number of variants identified across the 31 individuals genome or exome sequenced, using SAMtools mpileup, was 12,183,143. Table 3.4 describes the step-wise filtering of these variants to the 11,959,582 variants used in the pVAAST analysis and the 96,220 used in the heuristic filtering analysis.

Table 3.4 Initial filtering of identified variants.

Number of variants	Step-wise filtering
12,183,143	All variants identified
11,959,582	Variants that are high quality in at least one sample (coverage depth ≥ 10, quality score ≥ 20), used for pVAAST analysis
4,085,033	Variants with alternate allele frequencies ≤ 1% in 1000GP Caucasian data (Phase 1 v3 April 2012 release)
134,370	Variants located in RefSeq (as at 12/07/2014) gene regions that were exonic, splicing, 3' and 5' untranslated regions, upstream, downstream and within ncRNA
96,220	Variants with alternate allele frequencies ≤ 1% in 1000GP Caucasian data (September 2014, Phase 3 v5) and the UK10K consortia data, used for heuristic filtering analysis

3.4.3 Analysis of LK0051 family using pVAAST

3.4.3.1 LK0051 pVAAST analysis (11,959,582 variants) using the default VAAST background and the 1000GP background from 100 Caucasian samples

In separate pVAAST analyses, the VAAST provided 1000GP cross-population based background and a self generated background from 100 Caucasian 1000GP samples were used with the LK0051 family target data, with three HM cases and two unaffected relatives, one of whom (LK0051-007) is an obligate carrier connecting two cases. Multiple genome-wide significant associations (VAAST genome-wide significance threshold: $P \leq 2.4\times10^{-6}$) were identified as shown in the Manhattan plots (Figure 3.4 A and C) however the Q-Q plots (Figure 3.4 B and D) show that the pVAAST observed −log P-values are much greater than the expected −log P-values indicating a Type 1 error inflation. This means the identified associations should not be considered further.

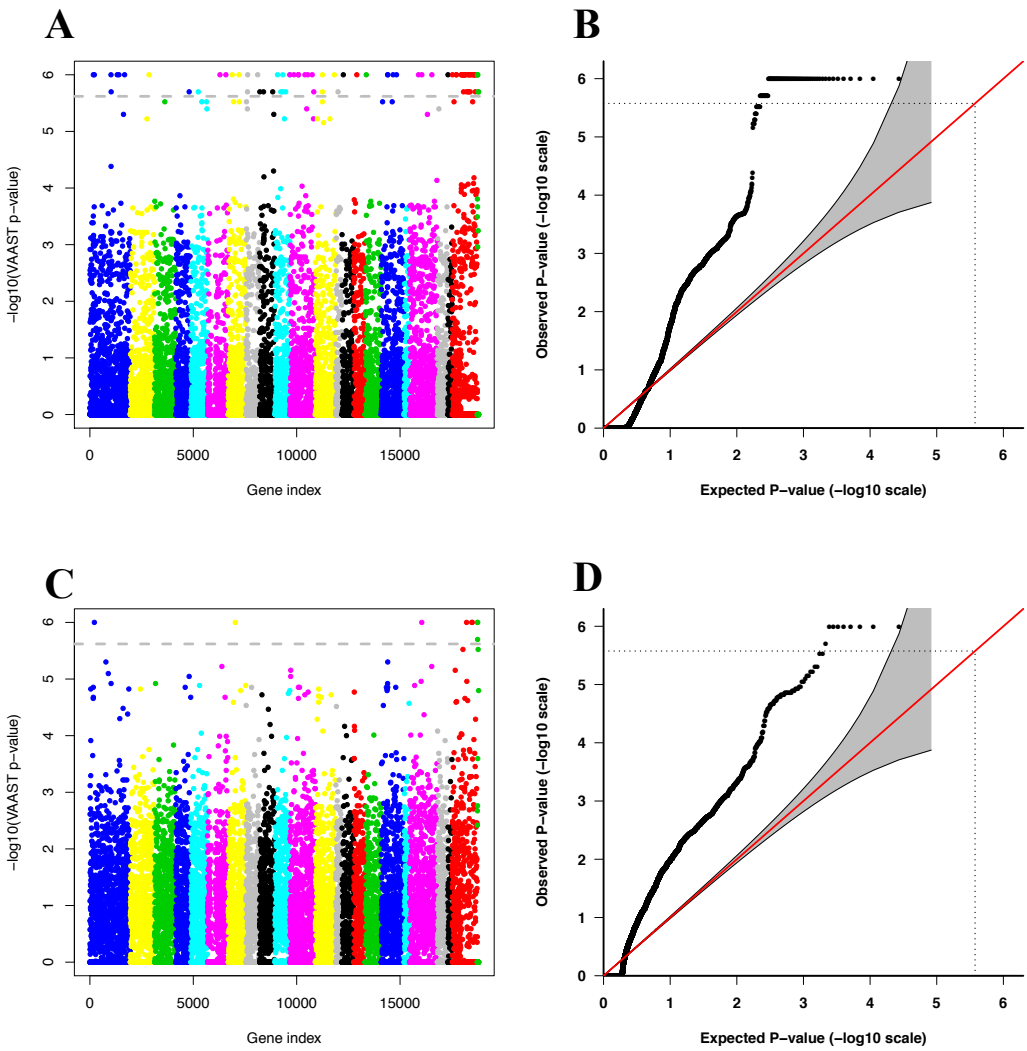


Figure 3.4 LK0051 pVAAST analysis results from using the default VAAST background and the 1000GP background from 100 Caucasian samples.
Manhattan plot (A, C) and Q-Q plot (B, D) results from the LK0051 pVAAST trial analysis using the VAAST developers sourced control background file (A and B) and a self generated control background file from 100 Caucasian samples from the 1000GP (C and D) that underwent the same variant calling methods as the target samples. The x-axis of the Manhattan plots show gene-based results (with VAAST gene index) in sequential chromosomes from chromosome 1 to chromosome Y. VAAST genome-wide significance threshold: $P \leq 2.4\times10^{-6}$ (dashed lines in figures A-D). Q-Q plot 95% confidence interval indicated by grey shading in B and D.

3.4.3.2 LK0051 target pVAAST analysis using a TFHMS sample background

The Type 1 errors resulting from use of the default VAAST background and the 1000GP background from 100 Caucasian samples in the pVAAST analysis with LK0051 indicated that a population, sequencing methodology, bioinformatics pipeline matched background with target and background samples variant called together was required, as per VAAST and pVAAST guidelines[254]. At the time of analysis to meet these guidelines with available data the decision was made to use the remainder of the TFHMS samples, N=26, combined together as a background control. Analysis with this background was run firstly with all 11,959,582 variants (Figure 3.5 A and B) and then with the 96,200 rare variants with 1000GP and UK10K Caucasian minor allele frequencies ≤1% (Figure 3.5 C and D). Both strategies reduced the inflated Type 1 error originally observed with the previous backgrounds. However the Q-Q plot when all 11,959,582 variants (Figure 3.5 B) still shows abnormal inflation of the observed −log P-values. As described in[254] VAAST-based analysis is a conservative test so the observed P-values are typically lower than the expected P-values when target and background are well matched (for population, sequencing methodology, bioinformatics pipeline) and sample sizes are small. The Q-Q plot when using 96,200 rare variants only (Figure 3.5 D) is representative of the typical Q-Q plot described in[254] so the results from this analysis can be considered further.

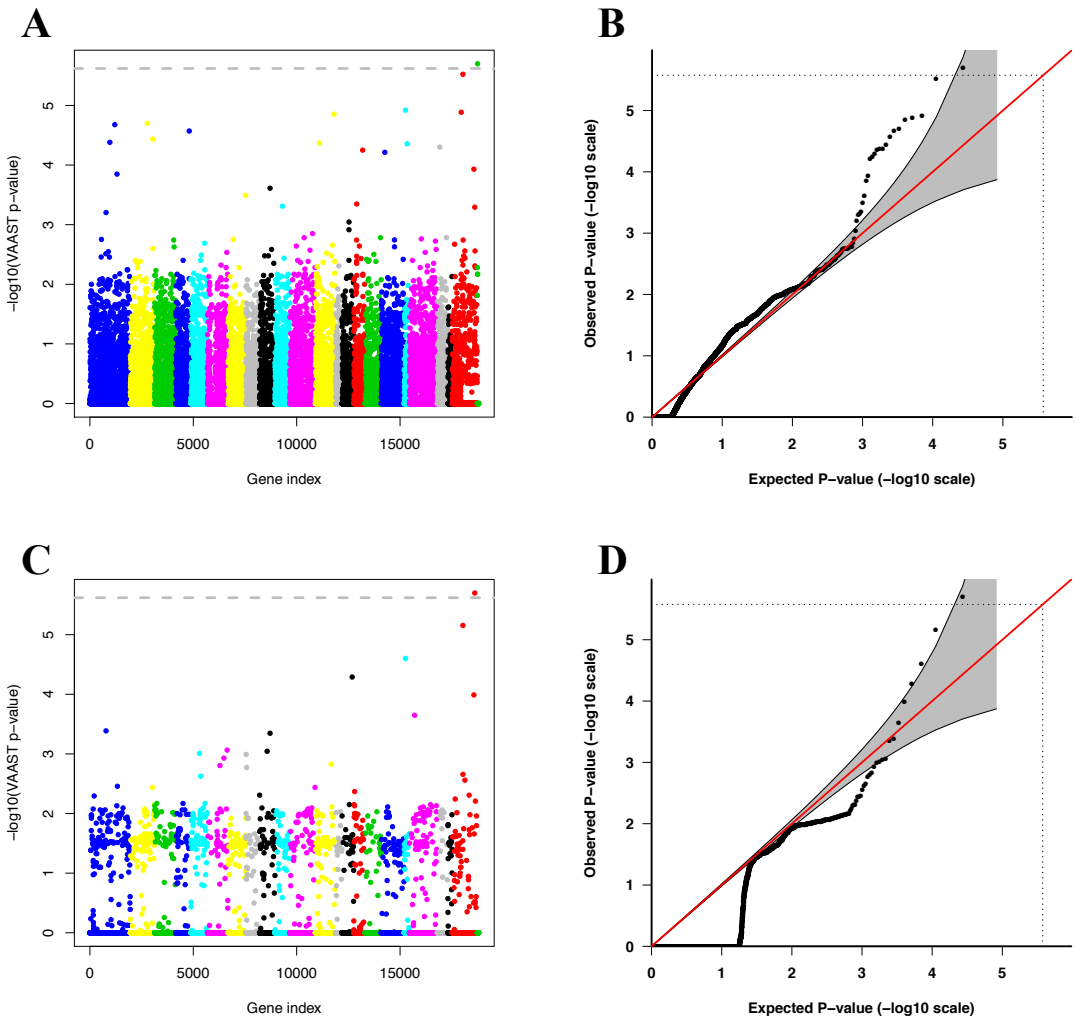


Figure 3.5 LK0051 pVAAST analysis results from using a TFHMS sample background.

Manhattan plot (A, C) and Q-Q plot (B, D) results from the LK0051 pVAAST trial analysis using the self generated control background file from the remaining 26 TFHMS samples that are from the same population and underwent an identical sequencing methodology and bioinformatics pipeline to identify variants. The x-axis of the Manhattan plots show gene-based results (with VAAST gene index) in sequential chromosomes from chromosome 1 to chromosome Y. VAAST genome-wide significance threshold: $P \leq 2.4\times10^{-6}$ (dashed lines in figures A-D). Q-Q plot 95% confidence interval indicated by grey shading in B and D.

From the pVAAST analysis restricted to rare variants, only one genome-wide significant result was obtained (Figure 3.4 C). The result of this analysis implicated the gene *MAGEC1* (melanoma antigen family C, 1) with a P-value of 1×10^{-6} listing four variants as contributing to this association. These variants are shown in Table 3.5. However all four variants are present in the unaffected non-obligate carrier family member LK0051-165 and there is no evidence of inheritance of this variant in the LK0051-001, LK0051-007 and LK0051-128 cluster as the unaffected obligate carrier, LK0051-007, is homozygous reference at all four variants.

Table 3.5 LK0051 pVAAST analysis trial results using rare variants and TFHMS background.

Location	Variant	LK0051-001	LK0051-007	LK0051-128	LK0051-159	LK0015-165
chrX:140993745	T>A	0/0	0/0	0/1	0/1	0/1
chrX:140993751	A>G	0/0	0/0	0/1	0/1	0/1
chrX:140993755	C>A	0/0	0/0	0/1	0/1	0/1
chrX:140993913	A>G	0/1	0/0	0/1	0/1	0/1

0/0 = homozygous reference for variant, 0/1 = heterozygous for variant

3.4.4 Analysis of TFHMS families using a heuristic tiered analysis strategy

The step-wise heuristic tiered analysis as outlined in section 3.3.7 was applied by using multiple Tier Two sharing strategies across the study families. For each family the results presented below show in table format the tier-by-tier results for each sharing analysis and a table of the final resulting variants for each family. All prioritised variants were heterozygous.

3.4.4.1 The final variant set for heuristic tiered analysis

After removing variants present in unaffected, non-obligate carrier relatives of HM cases across the five study families 44,692 variants remain for analysis. As shown in the Venn diagram in Figure 3.6, not all families carry each of the 44,692 variants so a smaller number of variants are applicable in each family in the heuristic tiered analysis.

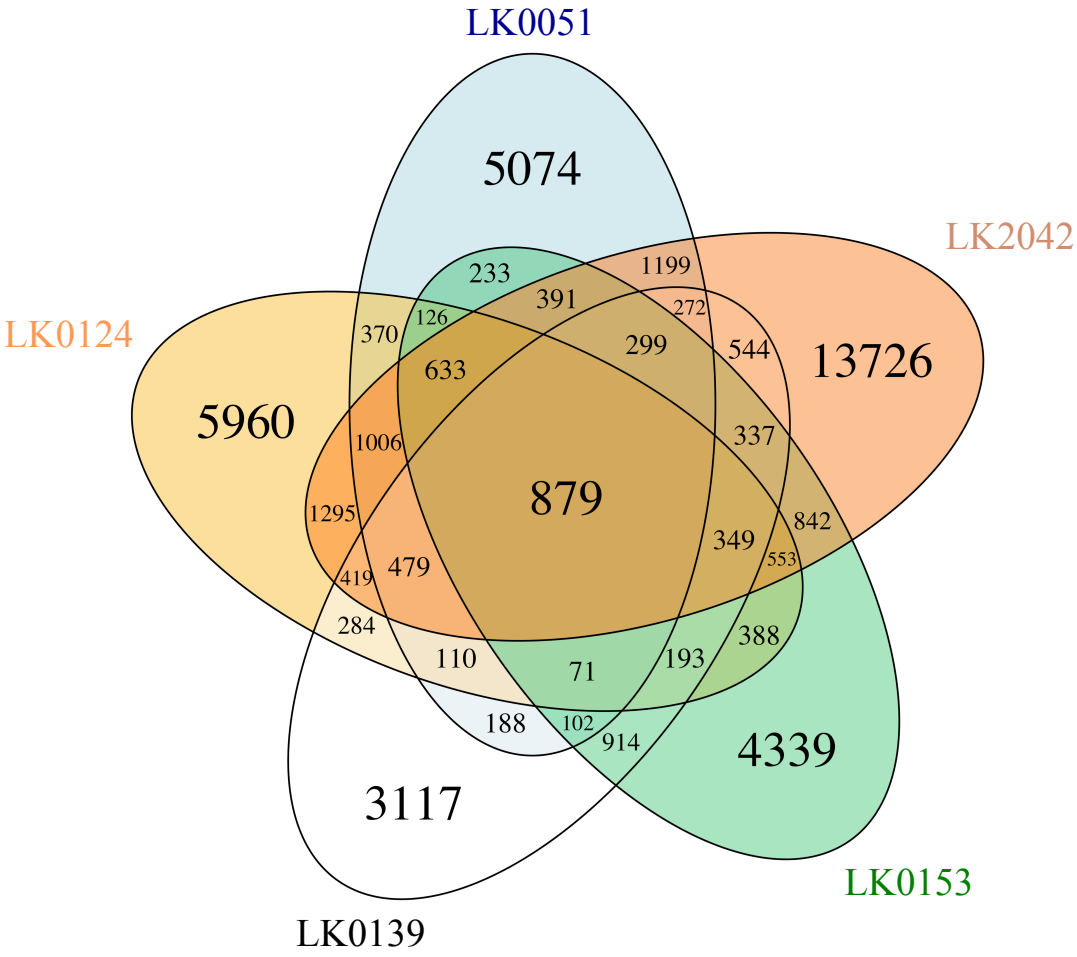


Figure 3.6 Venn diagram showing the sharing of 44,692 rare genetic variants across the five TFHMS families sequenced.

3.4.4.2 Heuristic tiered analysis of the LK0051 family

Three family sharing Tier Two strategies were used in the analysis of the LK0051 family as shown in Figure 3.7.

The first strategy used was to identify variants shared between all three cases, LK0051-001, LK0051-128 and LK0051-159 and the obligate carrier, LK0051-007. This strategy hypothesises that the two closely related cases (LK0051-001, LK0051-128) share susceptibility variants with the distantly related case (LK0051-159). LK0051-007 is considered an unaffected obligate carrier of the disease susceptibility variants shared between LK0051-001 and LK0051-128.

The second strategy used was to identify variants shared between the closely related cases, LK0051-001 and LK0051-128 and the obligate carrier, LK0051-007. This strategy hypothesises that closely related cases will share susceptibility variants and that LK0051-159 does not carry the susceptibility variants due to their genetic distance (> 8 meioses).

The third strategy used was to identify non-shared variants present individually in each HM case. This strategy hypothesises that each HM case has their own genetic variants that in addition to the shared genetic predisposition contributed to HM development.

The results of these three strategies are shown in Table 3.6. From the tiered prioritisation analysis in LK0051 seven Tier Six variants were identified for laboratory-based follow-up as described in Table 3.7.

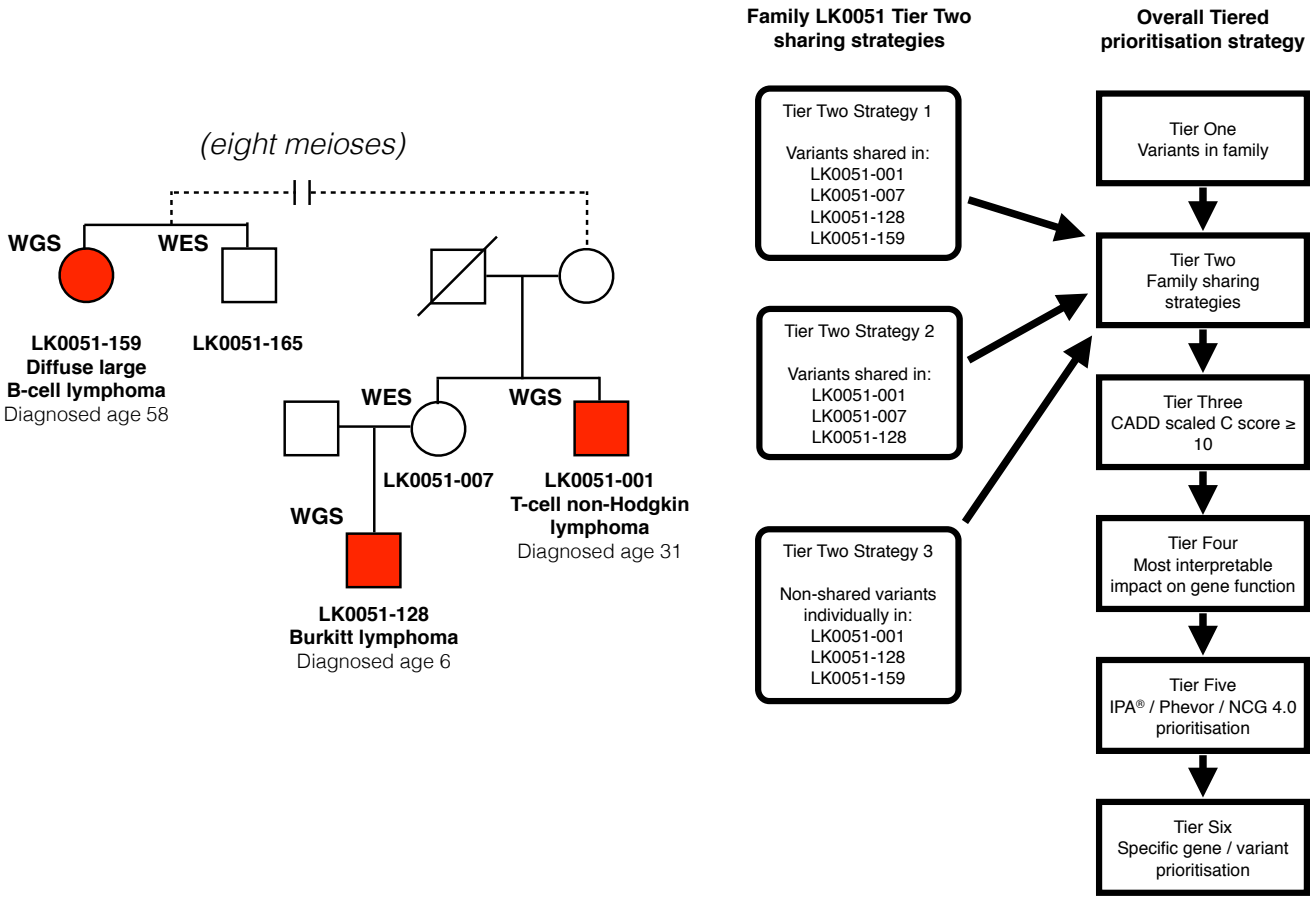


Figure 3.7 Schematic of the prioritisation strategies used in the analysis of family LK0051.
A reduced pedigree is shown for reference. The full pedigree is located in Figure 2.2 on page 51.

Table 3.6 LK0051 family heuristic-based analysis results.

	LK0051 Tier Two Sharing Strategies				
Prioritisation tiers	Strategy 1	Strategy 2	Strategy 3: LK0051-001	Strategy 3: LK0051-128	Strategy 3: LK0051-159
Variants in family	11,432				
Sharing strategy	995	1,755	3,201	2,617	3,029
CADD scaled C score ≥ 10	59	142	366	335	417
Most interpretable impact on gene function	3	28	62	56	61
IPA®/Phevor/NCG 4.0 prioritisation	3	8	27	26	29
Specific gene / variant prioritisation	2	5	--	--	--

Table 3.7 LK0051 family prioritised variants.

Family Strategy	Gene	Variant / dbSNP138 ID if known	Type	Functional effect	RefSeq Gene[237] biological function summary
S1	*NOTCH1*	chr9:139417464 T>G	Exonic SNV	T194P	Has a role in a variety of developmental processes by controlling cell fate decisions.
S1	*NF1*	chr17:29508805 T>G rs200962248	Splicing SNV	Changes the second intronic nucleotide immediately 3' after exon 7, could affect splicing efficiency	Negative regulation of Ras signal transduction pathway, mutations are linked to neurofibromatosis type 1, Watson syndrome and JMML.
S2	*TNFSF9*	chr19:6534728 G>C rs61750000	Exonic SNV	G139A	Tumour necrosis factor (TNF) ligand family member 9, a costimulatory receptor molecule in T lymphocytes, also involved in antigen presentation process in B lymphocytes.
S2	*TDP2*	chr6:24658126 C>T rs200729372	Exonic SNV	S144N	Member of a superfamily of divalent cation-dependent phosphodiesterases. It associates with CD40, TNF receptor-75 and TNF receptor associated factors. It inhibits nuclear factor-kappa-B activation. Sequence and structural similarities to endonucleases involved in DNA repair and activation of transcription factors.
S2	*MMP8*	chr11:102585288 T>G rs138686754	Exonic SNV	Synonymous variant, R397R changes codon bias	Is involved in extracellular matrix breakdown, stored in secondary granules in neutrophils and is able to degrade type I, II and III collagens.
S2	*LRP5*	chr11:68181292 C>A	Exonic SNV	T880N	A lipoprotein receptor that binds and internalizes ligands in receptor-mediated endocytosis, transduces signals by Wnt proteins and plays a key role in skeletal homeostasis.
S2	*PEX6*	chr6:42934551 G>A	Exonic SNV	R644W	Has a direct role in peroxisomal protein import and is required for PTS1 (peroxisomal targeting signal 1) receptor activity.

3.4.4.3 Heuristic tiered analysis of the LK0124 family

Two family sharing Tier Two strategies were used in the analysis of the LK0124 family as shown in Figure 3.7.

The first strategy used was to identify variants shared between at least two of the three cases LK0124-117, LK0124-179 and LK0124-202. With at least eight meioses separating each HM case, this strategy hypothesises that distantly related HM cases will share susceptibility variants.

The second strategy used was to identify non-shared variants present individually in each HM case. This strategy hypothesises that each HM case has their own genetic variants that in addition to the shared genetic predisposition contributed to HM development.

The results of these two strategies are shown in Table 3.8. From the tiered prioritisation analysis in LK0124 three Tier Six variants were identified for laboratory-based follow-up as described in Table 3.9.

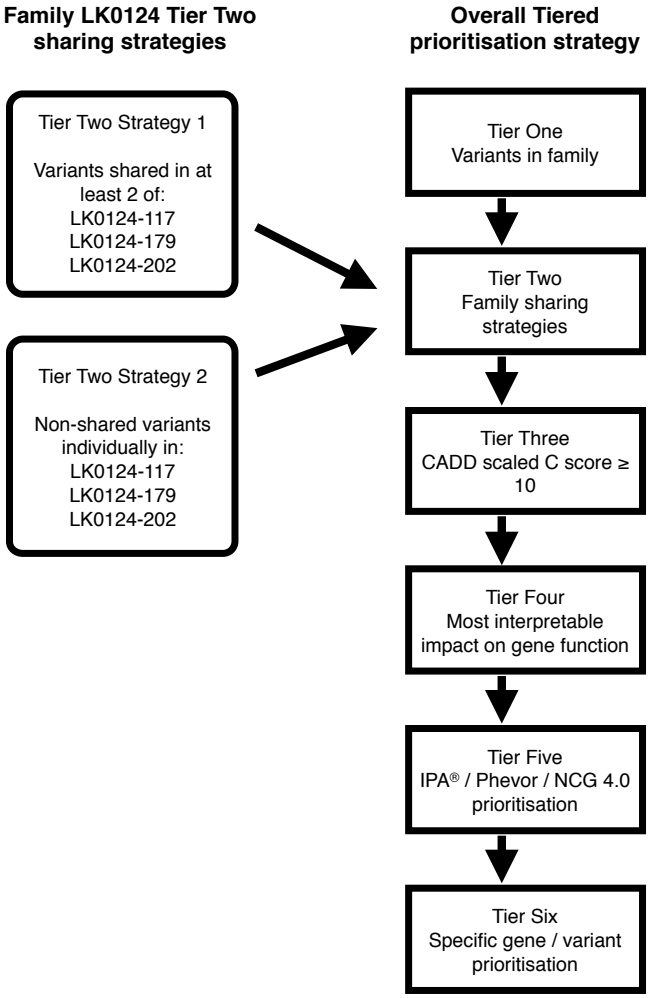


Figure 3.8 Schematic of the prioritisation strategies used in the analysis of family LK0124.
As the three individuals are distantly related a reduced pedigree is not an informative reference. The full pedigree is shown in Figure 2.3 on page 53.

Table 3.8 LK0124 family heuristic-based analysis results.

LK0124 Tier Two Sharing Strategies				
Prioritisation Tiers	Strategy 1	Strategy 2: LK0124-117	Strategy 2: LK0124-179	Strategy 2: LK0124-202
Variants in family	13,115			
Sharing strategy	2,457	3,680	3,530	3,430
CADD scaled C score ≥ 10	280	607	590	561
Most interpretable impact on gene function	40	85	111	89
IPA®/Phevor/NCG 4.0 prioritisation	25	38	50	39
Specific gene / variant prioritisation	3	--	--	--

Table 3.9 LK0124 family prioritised variants.

Family Strategy	Gene	Variant / dbSNP138 ID if known	Type	Functional effect	RefSeq Gene[237] biological function summary
S1	*NOTCH1*	chr9:139417464 T>G	Exonic SNV	T194P	Has a role in a variety of developmental processes by controlling cell fate decisions.
S1	*NF1*	chr17:29508805 T>G rs200962248	Splicing SNV	Changes the second intronic nucleotide immediately 3' after exon 7, could affect splicing efficiency	Negative regulation of Ras signal transduction pathway, mutations are linked to neurofibromatosis type 1, Watson syndrome and JMML.
S1	*STT3B*	chr3:31659458 T>G rs199778452	Exonic SNV	F384V	Is the catalytic subunit of a protein complex that transfers oligosaccharides onto asparagine residues.

3.4.4.4 Heuristic tiered analysis of the LK0139 family

Two family sharing Tier Two strategies were used in the analysis of the LK0139 family as shown in Figure 3.9.

The first strategy used was to identify variants shared between the two cases, LK0139-001 and LK0139-005. This strategy hypothesises that the parent-offspring affected pair share disease susceptibility variants.

The second strategy used was to identify non-shared variants present individually in each HM case. This strategy hypothesises that each HM case has their own genetic variants that in addition to the shared genetic predisposition contributed to HM development.

The results of these two strategies are shown in Table 3.10. From the tiered prioritisation analysis in LK0139 five Tier Six variants were identified for laboratory-based follow-up as described in Table 3.11.

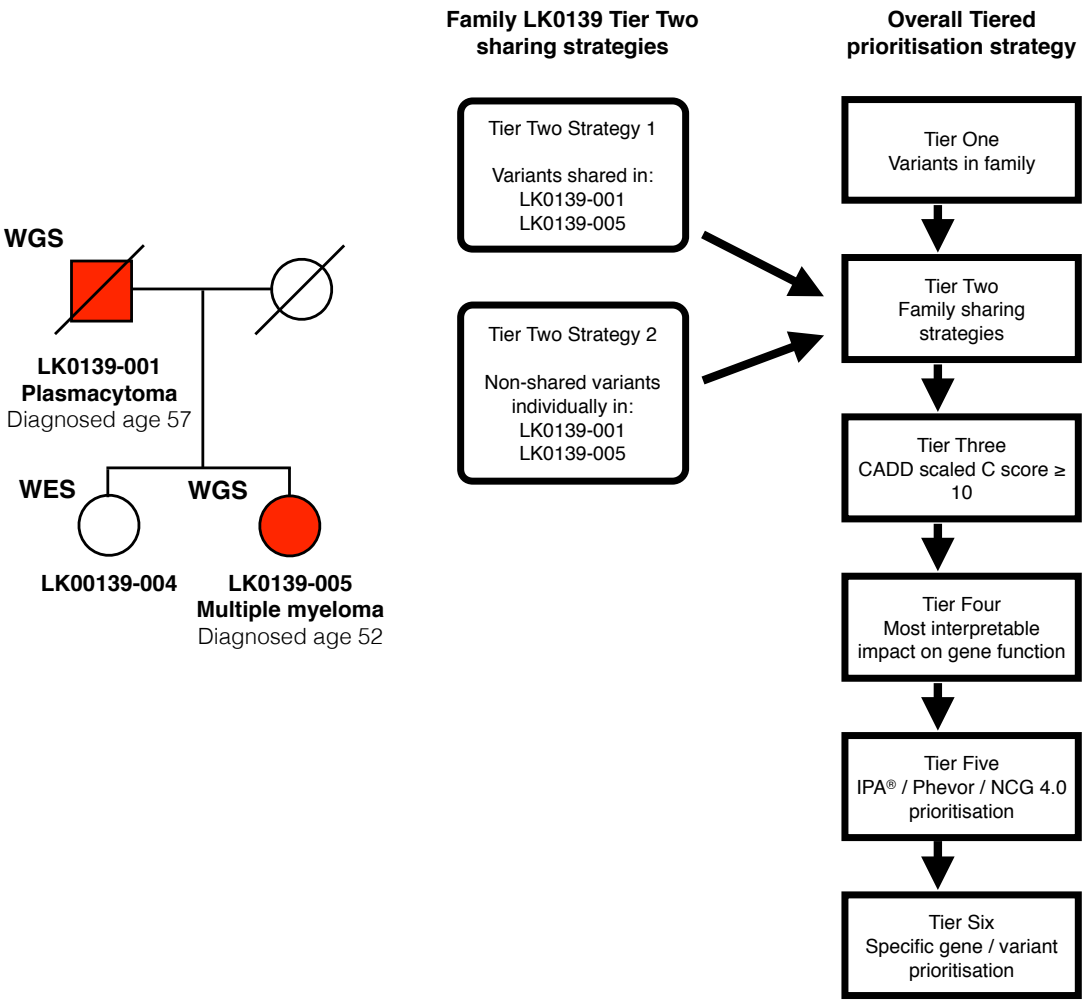


Figure 3.9 Schematic of the prioritisation strategies used in the analysis of family LK0139.
A reduced pedigree is shown for reference. The full pedigree is located in Figure 2.4 on page 55.

Table 3.10 LK0139 family heuristic-based analysis results.

LK0139 Tier Two Sharing Strategies			
Prioritisation Tiers	Strategy 1	Strategy 2: LK0139-001	Strategy 2: LK0139-005
Variants in family	8,557		
Sharing strategy	2,597	2,826	3,134
CADD scaled C score ≥ 10	387	336	439
Most interpretable impact on gene function	25	33	43
IPA®/Phevor/NCG 4.0 prioritisation	8	18	19
Specific gene / variant prioritisation	5	--	--

Table 3.11 LK0139 family prioritised variants

Family Strategy	Gene	Variant / dbSNP138 ID if known	Type	Functional effect	RefSeq Gene[237] biological function summary
S1	*NOTCH1*	chr9:139417464 T>G	Exonic SNV	T194P	Has a role in a variety of developmental processes by controlling cell fate decisions.
S1	*PABPC1*	chr8:101721839 C>A rs202074479	Exonic SNV	V365L	Is a poly(A) binding protein, that binds to 3' poly(A) tail of mRNA promoting ribosome recruitment and translation initiation. Also required for poly(A) shortening at the beginning of mRNA decay.
S1	*DUSP10*	chr1:221874839 TG>T	3'UTR Deletion	Predicted miRNA binding site for miR-129/129-5p miR-450a/450a-5p	Has a role in cell proliferation and differentiation through negative regulation of members of the MAP kinase superfamily, specifically p38 and SAPK/JNK.
S1	*PDE4DIP*	chr1:144863438 G>T rs140993521	Exonic SNV	Q1989K	Anchors phosphodiesterase 4D to the Golgi/centrosome region of the cell. Defects in gene may cause a myeloproliferative disease associated with eosinophilia.
S1	*SPHK2*	chr19:49132198 A>C rs200347384	Exonic SNV	Y378S	Catalyses phosphorylation of sphingosine to sphingosine 1-phosphate which mediates diverse cellular processes including migration, proliferation and apoptosis. In cancer can promote angiogenesis and tumourigenesis.

3.4.4.5 Heuristic tiered analysis of the LK0153 family

Three family sharing Tier Two strategies were used in the analysis of the LK0153 family as shown in Figure 3.10.

The first strategy used was to identify variants shared between the two HM cases, siblings LK0153-003 and LK0153-004, and their unaffected parent, LK0153-029, who is the genetic connection into the other HM affected members of this family. This strategy hypothesises that closely related cases share susceptibility variants and that in this context LK0153-029 is an unaffected obligate carrier of these variants.

The second strategy used was to identify variants shared between the two HM cases, LK0153-003 and LK0153-004, which were not present in their unaffected parent, LK0153-029. This strategy hypothesises that the siblings LK0153-003 and LK0153-004 share HM susceptibility variants that were not inherited from their parent LK0153-029.

The third strategy used was to identify non-shared variants present individually in each HM case. This strategy hypothesises that each HM case has their own genetic variants that in addition to the shared genetic predisposition contributed to HM development.

The results of these three strategies are shown in Table 3.12. From the tiered prioritisation analysis in LK0153 seven Tier Six variants were identified for laboratory-based follow-up as described in Table 3.13.

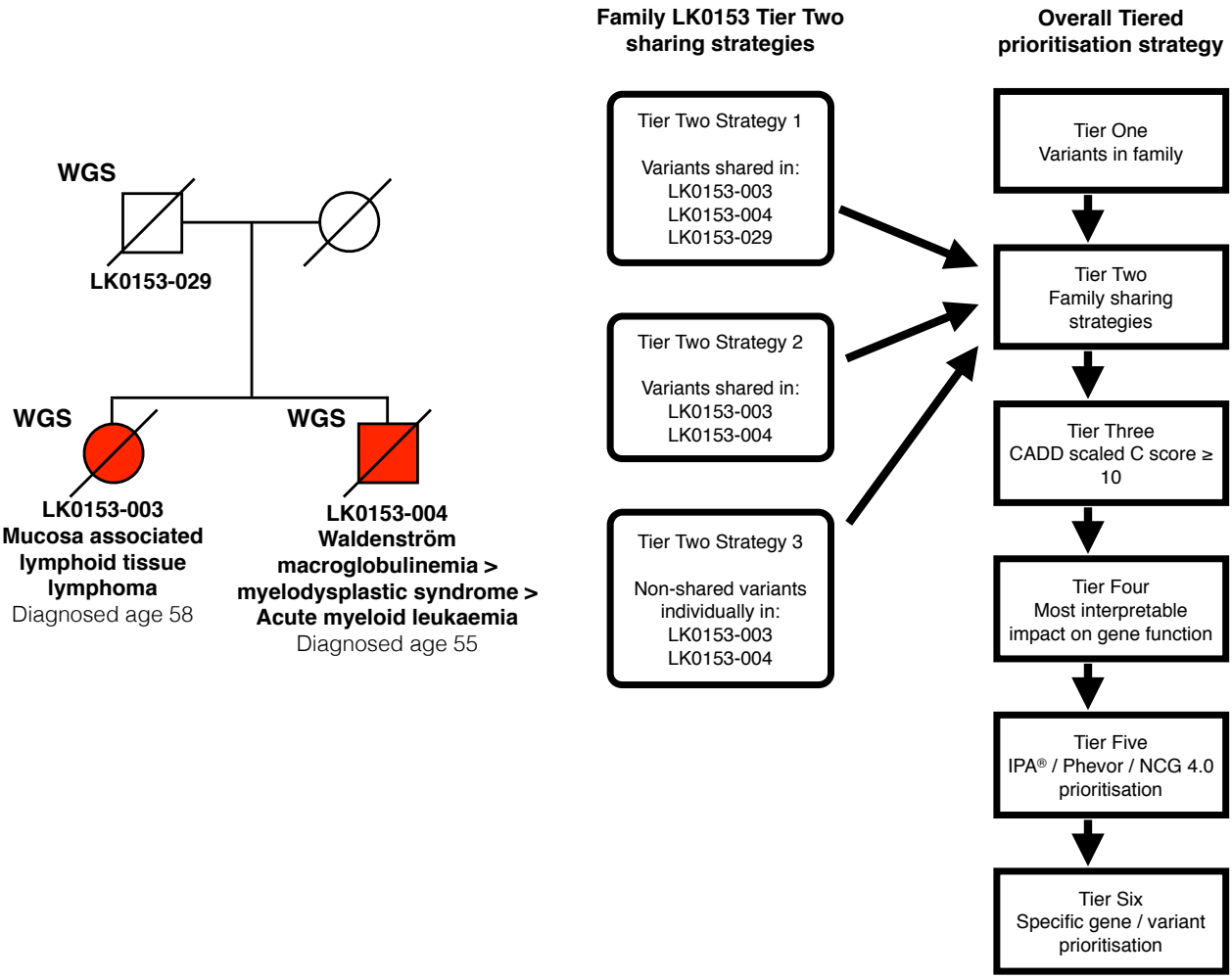


Figure 3.10 Schematic of the prioritisation strategies used in the analysis of family LK0153.
A reduced pedigree is shown for reference. The full pedigree is located in Figure 2.5 on page 57.

Table 3.12 LK0153 family heuristic-based analysis results.

LK0153 Tier Two Sharing Strategies				
Prioritisation Tiers	Strategy 1	Strategy 2	Strategy 3: LK0153-003	Strategy 3: LK0153-004
Variants in family	10,649			
Sharing strategy	1,388	1,218	1,705	1,711
CADD scaled C score ≥ 10	191	215	253	210
Most interpretable impact on gene function	32	23	25	26
IPA®/Phevor/NCG 4.0 prioritisation	11	12	14	13
Specific gene / variant prioritisation	3	4	--	--

Table 3.13 LK0153 family prioritised variants.

Family Strategy	Gene	Variant / dbSNP138 ID if known	Type	Functional effect	RefSeq Gene[237] biological function summary
S1	HAL	chr12:96371767 A>G rs150591434	Exonic SNV	W537R	Catalyses first reation in histidine catabolism.
S1	*RARS*	chr5:167929060 CCTT>C	Exonic deletion	F337del	Catalyses the aminoacylation of transfer-RNA by their congnate amino acid, linking amino acids with transfer-RNA codons.
S1	*RIPK2*	chr8:90796371 GTAA>G	Splicing deletion	Deletes the second to fifth intronic nucleotides immediately 3' after exon 8, could affect splicing efficiency	Member of the receptor-interacting protein family of serine/threonine protein kinases. A component of complexes in the innate and adaptive immune pathways, activator of NF-kappaB and inducer of apoptosis.
S2	*GIT1*	chr17:27904190 G>A rs202085570	Exonic SNV	R353W	No summary reported. Literature indicates a role in cancer cell migration and metastasis.
S2	*MET*	chr7:116381047 A>G rs374733251	Exonic SNV	T557A	A proto-oncogene associated with papillary renal carcinoma, is the hepatocyte growth factor receptor and encodes tyrosine-kinase activity.
S2	*NAT10*	chr11:34152973 C>A rs146685334	Exonic SNV	A472D	No summary reported. Literature indicates a role in cytokinesis regulation.
S2	*NID2*	chr14:52509033 G>A rs143412278	Exonic SNV	H539Y	Cell-adhesion protein that binds collagens I, IV and laminin and may be involved in basement membrane structure maintenance.

3.4.4.6 Heuristic tiered analysis of the LK2042 family

Two family sharing Tier Two strategies were used in the analysis of the LK2042 family as shown in Figure 3.11.

The first strategy used was to identify variants shared in two or more distantly related cases or obligate carriers. This strategy hypothesises that distantly related (> 8 meioses) cases and obligate carriers in a large family with multiple occurrences of HMs will share disease susceptibility variants.

The second strategy used was to identify non-shared variants present individually in each HM case. This strategy hypothesises that each HM case has their own genetic variants that in addition to the shared genetic predisposition contributed to HM development.

The results of these two strategies are shown in Table 3.14. From the tiered prioritisation analysis in LK2042 ten Tier Six variants were identified for laboratory-based follow-up as described in Table 3.15.

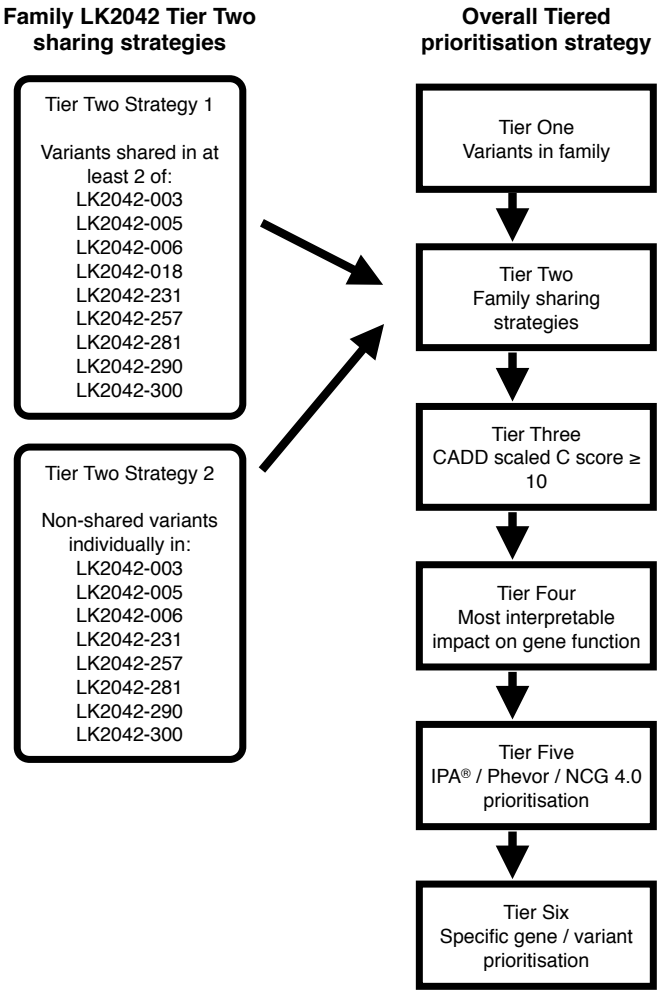


Figure 3.11 Schematic of the prioritisation strategies used in the analysis of family LK2042.

As the sequenced individuals are distantly related a reduced pedigree is not an informative reference, the full pedigree is shown in Figure 2.6 on page 59.

Table 3.14 LK2042 family heuristic based analysis results.

	LK2042 Tier Two Sharing Strategies								
Prioritisation Tiers	S1	S2: LK2042-003	S2: LK2042-005	S2: LK2042-006	S2: LK2042-231	S2: LK2042-257	S2: LK2042-281	S2: LK2042-290	S2: LK2042-300
Variants in family	23,223								
Sharing strategy	5,687	2,502	1,104	2,460	1,821	1,670	2,205	2,523	929
CADD scaled C score ≥ 10	593	486	383	434	293	212	427	490	318
Most interpretable impact on gene function	96	89	86	67	43	25	78	92	86
IPA/Phevor/NCG 4.0 prioritisation	51	49	46	31	25	11	32	35	39
Specific gene / variant prioritisation	10	--	--	--	--	--	--	--	--

Table 3.15 LK2042 family prioritised variants.

Family Strategy	Gene	Variant / dbSNP 138 ID if known	Type	Functional effect	RefSeq Gene[237] biological function summary	Number shared in
S1	*NOTCH1*	chr9:139417464 T>G	Exonic SNV	T194P	Has a role in a variety of developmental processes by controlling cell fate decisions.	6
S1	*NF1*	chr17:29508805 T>G rs200962248	Splicing SNV	Changes the second intronic nucleotide immediately 3' after exon 7, could affect splicing efficiency	Negative regulation of Ras signal transduction pathway, mutations are linked to neurofibromatosis type 1, Watson syndrome and JMML.	7
S1	*PABPC1*	chr8:101721839 C>A rs202074479	Exonic SNV	V365L	Is a poly(A) binding protein, that binds to 3' poly(A) tail of mRNA promoting ribosome recruitment and translation initiation. Also required for poly(A) shortening at the beginning of mRNA decay.	3
S1	*B4GALNT4*	chr11:377329 T>G	Splicing SNV	Changes the second intronic nucleotide immediately 3' after exon 14, could affect splicing efficiency	No summary reported. Gene Ontology data indicates the protein has acetylgalactosaminyltransferase activity and is involved in metabolic processes.	4
S1	*L3MBTL2*	chr22:41615519 G>C	Exonic SNV	A233P	No summary reported. Gene Ontology data indicates the protein is involved in chromatin modification.	3
S1	*PRIM2*	chr6:57246878 T>C	Exonic SNV	L202S	Encodes DNA primase, an enzyme with a key role in DNA replication. Synthesises the small RNA primers used in cell to create Okazaki fragments in lagging DNA strand in DNA replications.	2
S1	*PRIM2*	chr6:57398207 G>T	Exonic SNV	Inframe STOP codon	As above.	3
S1	*SAC3D1*	chr11:64811900 C>T	Exonic SNV	R260C	No summary reported. Gene Ontology data indicates that protein is involved in mitotic nuclear division.	2
S1	*SYVN1*	chr11:64898255 T>G	Exonic SNV	T328P	Encodes a protein involved in endoplasmic reticulum (ER) protein degradation of unfolded proteins that accumulate during ER stress.	3
S1	*TBX21*	chr17:45820022 A>C	Exonic SNV	T180P	Encodes a transcription factor involved in controlling the expression of the hallmark cytokine on T-cells and natural killer cells, interferon-gamma.	3

3.5 Discussion

3.5.1 pVAAST probabilistic analysis

The pVAAST probabilistic approach to disease gene identification has been successful previously in other pedigree-based contexts such as a family with dominant inheritance of cardiac septal defects[223] and in familial breast cancer and lynch-syndrome spectrum cancers[255]. It is applicable to inherited conditions, such as familial HMs, that have disease penetrance complexities. However when the pVAAST approach was trialled with data from the LK0051 family the results indicated further development and an alternative approach was required. Quantile-quantile plots of the results from the analyses using the default VAAST background and a background created from 100 Caucasian samples showed striking inflation of the observed −log P-values indicating a Type 1 error (Figure 3.4 B and D). In the context of VAAST and pVAAST this means that the target data and background control data are not matched.

The default VAAST background sourced from the VAAST developers is built from variants called from across the 1000GP populations, so firstly as was expected there was a ethnicity / population mismatch between the multi-population background and the LK0051 family Caucasian samples used in the target file. Additionally the variants used to build the default VAAST background were not identified using the same bioinformatics pipeline as the target LK0051 samples, which can also cause Type 1 error inflation [254]. For the background created from the 100 Caucasian samples, 70 Caucasian exomes and 30 Caucasian genomes from the 1000GP were downloaded and variants were called using the same methodology as the LK0051 target samples. Compute cluster storage restrictions were the limiting factor in the number of Caucasian samples accessed in the 1000GP. It was expected that using the same variant calling method as the target samples would improve the Type 1 error because of the matched variant calling pipeline to the target samples reducing biases from using a different variant calling software. Some improvement was seen (Figure 3.4 D vs. B) with the right tail of the Q-Q plot moving closer to the expected distribution, however observed −log P-values were still inflated. Given that the samples are now population ethnicity matched, Caucasian 1000GP samples and Caucasian Tasmanian study samples, this suggests that the inflation is either caused by continued population stratification with Tasmanian Caucasians being genetically

different to the 1000GP Caucasians, which is unlikely, or the effect is from differences in the sequencing platform and bioinformatics alignment and variant calling pipelines. The 1000GP samples had been aligned to the NCBI GrCh37 reference genome, whereas the TFHMS LK0051 samples were aligned to UCSC's hg19 reference genomes. Differences in the encoding of the two versions of the reference genome prevented multi-sample variant calling of TFHMS and 1000GP samples together. This is the most likely explanation for the Type 1 error inflation as VAAST and pVAAST protocols advise strict target and background matching. The 1000GP genome and exome BAM files should be demultiplexed to FASTQ raw read files and realigned to the UCSC hg19 following the same bioinformatics pipeline as the TFHMS samples and then target and background samples multi-called together using the same variant calling software. This approach should control completely for the Type 1 error inflation. This was not, however, feasible with available computational resources and time at the time of analysis. Thus the results from the LK0051 pVAAST analysis using these two backgrounds were not further pursued at this time. Rigorously matched controls, ideally Tasmanian population matched, will be required as will further trialling of different analysis parameters in the pVAAST software dependant upon the disease family.

To control for the factors influencing the Type 1 error in the analyses using the first two backgrounds a third background control file to analyse LK0051 against was formed from the remaining 26 TFHMS samples, which had all undergone identical bioinformatics pipeline processing and had been variant called together. Initial results from this background were promising with a decrease in Type 1 error inflation seen (Figure 3.5 B) however the Q-Q plot was still abnormal in comparison to what is expected for the conservative VAAST-based analysis[254]. A supplementary pVAAST analysis by filtering the 11,959,582 variants to 96,200 rare variants with MAF $\leq$ 1% in the 1000GP and UK10K Caucasian populations improved the Q-Q plot, meeting the expectations outlined in Kennedy *et al.*[254] (Figure 3.5 D). The result of this analysis using only rare variants identified one genome-wide significant hit, *MAGEC1* on chromosome X with a P-value of 1×10^{-6}. However analysis of the variants contributing to this association revealed that they were not consistent with the expected inheritance pattern due to absence of the variants in the obligate carrier LK0051-007. Additionally, LK0051-128 was chosen to be the representative disease

case for the family, thus requiring the variant to be present in LK0051-128 and the default options used in the analysis applied a penalty to incomplete penetrance, so any variants inherited in LK0051-001, LK0051-007 and LK0051-128 would have been statistically penalised because LK0051-007 is not affected. Analysis removing this penalty did not alter the findings.

3.5.2 Heuristic filtering-based analysis using a tiered approach

The probabilistic approach, when successful, is ideally complemented by a heuristic approach. Such an approach is logical and hypothesis driven, relying on a defined series of questions to filter the total number of variants identified to a smaller number, depending upon the questions asked. With further work required on the application of pVAAST, a heuristic filtering-based approach to variant prioritisation was adopted as the main analysis method in this study.

The heuristic approach used here was developed to focus on variants shared between related cases, with sharing analysis strategies constructed for each family. Then by using ANNOVAR, and several other variant-based annotation tools (as detailed in Table 3.2) variants were prioritised using a tiered strategy to find the shared variants most likely to contribute to disease development in each of these families.

The use of WES of individuals in these families supplemented the WGS information and aided the prioritisation strategy employed. However as shown in Figure 3.3, overall coverage of exome capture regions was lower than WGS, with approximately 20% of the WES regions across samples having a read depth lower than the threshold of ≥10 used in this analysis. This suggests that a number of informative variants in these individuals have been missed and is a limitation of the use of a combined WES and WGS approach in these families.

3.5.2.1 Variants prioritised in family LK0051

In family LK0051 two sharing strategies were applied in the tiered prioritisation. The first was to examine variants shared between all three HM cases and the obligate carrier, the second was to focus on variants in the uncle-nephew case pair and their connecting female relative, separating them from their affected distant cousin. These

two strategies together identified seven variants of interest for further laboratory-based follow-up (Table 3.7).

In the second sharing strategy in this family, two variants in *NEB* and *OBSCN* were specifically excluded in Tier Five due to their identification in the NCG 4.0 database[249] as likely false positive results due to gene length[249]. From Tier Five to Tier Six a further three variants were excluded after closer examination of the specific variant effects. Two of these were excluded as they were predicted to be non-conserved residues by Project HOPE[250]. That is, the presence of the variant amino acid in other homologous sequences for that protein, indicates that the change is unlikely to affect protein function. The third was a predicted splicing variant which was excluded based this variant being unlikely to affect the critical splicing nucleotides at the exon-intron junction, as reviewed in Baralle and Baralle[251].

Of the variants prioritised in LK0051, both the *NF1* and *NOTCH1* variants were also present in the analyses of two (LK0124, LK2042) and three (LK0124, LK0139, LK2042) other sequenced families respectively. *NOTCH1* in particular has been shown to be somatically mutated in HMs and has a known role in haematopoietic cell differentiation[177,178,184,256-260]. Mutation to *NF1* is known to cause Neurofibromatosis type 1 (NF1) (MIM #162200) a familial cancer syndrome, of which development of a HM, juvenile myelomonocytic leukaemia, can be a feature amongst other diverse phenotypes. Occurrence of NF1 is not suspected in the families carrying the mutation based on known medical histories, however NF1 has been shown to have diverse presentations in different families[261]. Evidence of the occurrence of HMs in *NF1* mutation carriers without other NF1 symptoms has also been reported[262,263].

The five variants arising from the second sharing analysis strategy in LK0051 in *TNFSF9*, *MMP8*, *TDP2*, *LRP5* and *PEX6* all present as interesting candidates for further laboratory-based consideration. A connection between HM predisposition and variants in these genes is a novel finding and disruption to the biological roles of each of these genes could be hypothesised to contribute to malignancy development. These variants occur in the uncle-nephew pair, with a T-cell and a B-cell HM respectively. *TNFSF9* is of particular interest as the glycine to alanine amino acid change at position 139 affects a completely conserved residue across all homologous protein

sequence alignments. This together with known roles of *TNFSF9* in both T-cell and B-cell biology strengthens interest in this gene and variant in HM predisposition.

3.5.2.2 Variants prioritised in family LK0124

In family LK0124, due to more distant relationships between the three HM cases sequenced, the sharing strategy chosen was to focus on variants shared in two or more cases. The tiered analysis identified three variants of interest, two are the same afore mentioned *NOTCH1* and *NF1* variants in family LK0051. The third is a SNV in *STT3B* causing the F384V amino acid change.

3.5.2.3 Variants prioritised in family LK0139

Family LK0139 presented an opportunity to identify shared variants between a father-daughter HM affected pair who had plasmacytoma (a multiple myeloma precursor condition) and multiple myeloma respectively. Five shared variants were identified including the previously mentioned variant in *NOTCH1*. Of the remaining four variants, one in *PABPC1* was also observed in the LK2042 family. *PABPC1* has an important role in mRNA translation and the amino acid change identified, V365L, may be altering this role. However, while the lysine residue has not been observed at this position, other amino acids with similar properties to lysine have been, which may suggest that this variant does not have a functional effect on the protein. A 3'UTR mutation in a TargetScanS predicted miRNA binding site in *DUSP10* was also identified in this family. Confirmation that miRNAs bind at this 3' location, and that this binding is disturbed by the variant would be required to show the functional role of this variant. The remaining two variants of interest in the LK0139 family are amino acid changes that occur in *PDE4DIP* (Q1989K) and *SPHK2* (Y378S). Both genes are known to have a role in cell proliferation, which may indicate a potential role in HM development when disrupted.

3.5.2.4 Variants prioritised in family LK0153

Sequencing in family LK0153 was conducted in an affected sibling pair and their unaffected father who had a family history of HMs. Two family sharing analyses were conducted to identify variants shared between the siblings, and separately

variants that the siblings also shared with their father, as it was possible that the disease susceptibility shared between the siblings was independent of their father. The two analyses identified seven variants, three between all family members and four shared between the siblings and not their father. Of these seven variants, a splicing variant in *RIPK2* that deletes three intronic nucleotides immediately after exon eight occurred in all three sequenced family members. This variant may produce a *RIPK2* transcript variant that affects the reported role of this gene in apoptosis, indeed previously a splice variant without exon 2 was identified which resulted in a loss of activity of the RIPK2 protein[264]. Shared between the affected siblings but not their father were four exonic variants resulting in amino acid changes in functional protein domains. Particularly of interest are the variants in *MET*, because it is a known proto-oncogene[265] and *NAT10*, because this gene has suggested roles in cytokinesis regulation destabilization of which can contribute to cancer[266].

3.5.2.5 Variants prioritised in family LK2042

Family LK2042, as for family LK0124, is a large family with the eight sequenced cases each being distantly related. This presented an analytical challenge, as identification of variants with Mendelian inheritance was not possible. As for LK0124 the sharing strategy selected for LK2042 was to focus on variants in two or more cases. This identified over fifty Tier Five variants for individual prioritisation. Individual analysis of each of these variants prioritised ten in Tier Six. Given the genetic distance between the related cases in this family, a shared variant analysis may not be the best analysis option in this family. However variants in interesting candidate genes were prioritised, including the previously mentioned *NOTCH1*, *NF1* and *PABPC1* variants. In addition two variants in the gene *PRIM2* encoding a component of the primase enzyme responsible for the synthesis of Okazaki fragments during DNA replication were identified[267]. *PRIM2* has been shown in a mouse model of virus induced HM to be a candidate gene involved in HM development[268]. In LK2042, one variant caused an amino acid change, L202S, with the lysine residue being conserved across 53% of alignments according to the Pfam database[246] and no evidence of the serine residue occurring in homologous sequences. This variant was identified in LK2042-281 and LK2042-300. The second *PRIM2* variant introduces an in frame stop codon in exon nine of the gene in LK2042-006, LK2042-290 and

LK2042-300. The occurrence of multiple variants in *PRIM2* in the LK2042 family warrants further investigation of this gene in HM development.

3.5.2.6 Non-shared variants prioritised in each HM case

Additionally for each affected HM case a secondary Tier Two analysis was undertaken to identify and prioritise non-shared variants. This analysis revealed that non-shared variants, those that are not shared between related cases and are just present within an individual HM affected family member, might also be relevant to disease predisposition. Such solo variants may work in concert with the identified shared candidate predisposition variants to contribute to the individual's risk of HM development. A group of sequenced Tasmanian population controls would assist in identifying whether individuals in these families are enriched for more Tier Five and Tier Six variants than would be expected in unrelated controls. This is planned, but was not available at the time of analysis.

Non-shared variants are not the focus of this dissertation, so accordingly a Tier Six variant-based analysis was not undertaken, but it is interesting to note that in these analyses individuals had, on average, 35 (95% C.I. = 27–43) Tier Five variants. Several of these variants are striking candidates for HM disease predisposition, including the *JAK2* V617F mutation in LK2042-290 who has essential thrombocythaemia (a MPN, a diagnostic feature of which is *JAK2* V617F mutations; in this individual this mutation was confirmed in their clinical record). While it is likely that the *JAK2* mutation was somatically acquired, the mother of LK2042-290 also had a MPN, which may indicate the inheritance of the *JAK2* mutation, or inheritance of other genetic variants contributing to a MPN predisposition.

3.6 Conclusion

This chapter detailed the bioinformatics approach to quality assessment and variant identification from the 31 next generation sequenced samples used in this study. Initial variant filtering and annotation, as described, was conducted primarily using ANNOVAR[226] and other web-based tools. This strategy reduced the total number of variants identified using SAMtools mpileup from 12,183,143 to 96,220 rare SNVs and indels in coding or potential regulatory regions. This set of variants formed the

basis of a family-based tiered prioritisation strategy to identify potential interesting variants for further follow up in relation to HM development. Future work for variant identification using the NGS data from these samples would be to explore the identification and prioritisation of structural variants, particularly copy number variants.

In total, twenty-six very promising novel candidate variants with clear biological links to HMs, that have not been previously described in relation to HM predisposition, have been identified through a family-based analysis in this study. Each of these variants requires laboratory validation to confirm the presence of the variant in the sequenced samples. In particular, the occurrence of rare variants across families raises the concern that these variants may be false positives arising from artefacts related to the next generation sequencing and analysis process.

Laboratory validation ideally uses a NGS independent method to verify sequence variants. Chapter 4 details the laboratory validation work completed on a selection of these variants, and in particular the high-throughput screening of two variants across the TFHMS resource and a group of Tasmanian population controls.

Chapter 4 - Validation and population screening of selected prioritised variants

4.1 Introduction

The prioritisation of sequence variants through bioinformatics analysis of NGS data is a strategy designed to identify potential disease susceptibility variants. Not all prioritised variants will be true sequence variants, or necessarily be truly disease causing. The occurrence of false positive sequence variants is an expectation in NGS data[269] and so a NGS independent method of variant detection, such as Sanger sequencing, is required for variant validation. The occurrence of prioritised variants in diseased individuals is not evidence that the variant truly contributes to disease. A prioritised and validated variant must be screened within an ethnicity matched control population to ensure that the variant is not common within that specific population, but rare within community control data such as the 1000GP and EVS populations. Additionally a prioritised, validated and rare variant should be screened, where samples are available, in the specific disease resource of individuals. This is a first step in determining whether the variant might contribute more widely to disease susceptibility.

In this study twenty-six variants were identified from the analyses of NGS data in five TFHMS families, as described in Chapter 3, that are promising candidates of HM susceptibility in these families. To first confirm that the identified variants are not artefacts of the NGS process, laboratory-based confirmation of each of these variants is required. Variant validation by Sanger sequencing was necessarily prioritised due to time and funding constraints and so was only conducted on the prioritised variants in families LK0051, LK0124, LK0139 and LK0153. From the successfully validated variants, two genes for which there was the greatest supporting biological evidence such that they could be hypothesised to be involved in HM development, were chosen from family LK0051 and high-throughput genotyping screens using a TaqMan genotyping assay were completed. Allele frequencies were used to test whether either variant was enriched in the TFHMS resource. Subsequently, based on the results of

this analysis, one of these genes was screened by amplicon-based NGS on a MiSeq system across 96 HM cases, in search of other mutations in this gene that may be contributing to HM development.

4.2 Aims

The objective of this analysis was to first use an independent method to validate prioritised sequence variants identified through NGS. Then, for selected validated variants, conduct a high-throughput genotyping screen of the TFHMS resource and a group of population controls to establish further evidence for the potential contribution of the variants to HM susceptibility.

4.3 Methods

4.3.1 Sanger sequencing confirmation of identified variants

4.3.1.1 Oligonucleotide DNA PCR primer design

Sanger sequencing was used to confirm variants identified through NGS. PCR amplicons were designed to amplify a 200-500 base pair fragment spanning the variant site using the UCSC genome browser to obtain the reference sequence in these regions as well as known SNPs for masking prior to primer design[234]. Primer-BLAST[270] was used to design sequence specific oligonucleotide DNA primers, using the human genome for primer pair specificity checking. Primer sequences are listed in Appendix 4.1.

Primers were synthesised by GeneWorks or Sigma Aldrich. Primers were optimised using gradient PCR to determine appropriate annealing temperatures for PCR amplification. Primers were used for both PCR amplification and Sanger sequencing.

4.3.1.2 DNA amplification by PCR

PCR amplicons were amplified from 10 ng/µL genomic DNA, obtained as per Section 2.2.4, according to the following reaction conditions:

10 µL reactions, per sample:

5 µL	Promega GoTaq Green master mix, or Bioline MyTaq HS master mix
0.8 µL	Forward oligonucleotide primer at 10 µM
0.8 µL	Reverse oligonucleotide primer at 10 µM
2.4 µL	H_2O
1 µL	DNA at 10 ng/µL

Thermal cycling conditions:

GoTaq Green PCRs		MyTaq PCRs	
95ºC	3 mins	95ºC	2 mins
40 cycles of:		40 cycles of:	
95ºC	1 min	95ºC	10 secs
XºC	30 secs	XºC	10 secs
72ºC	1 min	72ºC	20 secs
72ºC	5 mins	4ºC	hold
4ºC	hold		

(X = optimised PCR annealing temperature)

Thermal cycling was conducted using Veriti thermal cyclers from Life Technologies. Each PCR was conducted with both a placenta derived human genomic DNA positive control (Bioline) and a DNA-free negative control. PCR products were visualised by gel electrophoresis (85 volts for 25 mins) on 1% agarose gels with SYBR Safe DNA gel stain (Life Technologies).

4.3.1.3 PCR product purification

10 µL PCR reaction products were prepared for sequencing by magnetic bead purification using Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer's instructions. 18 µL of beads were used for 10 µL PCR reactions. Purified PCR products were eluted from beads in 40 µL of nuclease free H_2O in preparation for Big Dye Terminator (BDT) PCR for Sanger sequencing.

4.3.1.4 BDT Sanger sequencing PCR

BDT PCR amplification was conducted using 1 µL of AMPure purified PCR product using either the forward or reverse primers from the initial PCR amplification according to the following reaction conditions:

10 µL reactions, per sample:

0.25 µL	BDT enzyme
1.75 µL	Sequencing buffer
1 µL	Primer at 3.3 µM
6 µL	H_2O
1 µL	AMPure purified PCR product

Thermal cycling conditions:

96ºC	1 mins
25 cycles of:	
96ºC	10 secs
50ºC	5 secs
60ºC	4 min
4ºC	hold

Veriti thermal cyclers from Life Technologies were used to conduct BDT PCR amplification.

4.3.1.5 BDT PCR product purification and capillary sequencing

Following PCR amplification the BDT PCR products were purified using Agencourt CleanSeq (Beckman Coulter) beads according to the manufacturer's instructions with 10 µL of beads used per 10 µL PCR reaction. Purified PCR products were eluted from beads with 40 µL of nuclease free H_2O. 30 µL of purified PCR product was loaded into an ABI 310 or ABI 3100 capillary sequencer. Sanger sequencing results were analysed using the Sequencher software package, version 4.10.1 from the Gene Codes Corporation.

4.3.1.6 TaqMan probe-based genotyping

TaqMan probe-based genotyping was used for a high-throughput assessment of selected variants. TaqMan genotyping probes were obtained from Life Technologies. Each selected variant required a custom probe design. These were designed according to the manufacturer's specifications. Briefly, up to 500bp upstream and downstream of the variant of interest was extracted from UCSC genome browser with masking of known SNPs and repeats (as determined by RepeatMasker) and submitted to Life Technologies. Variants were screened in all available TFHMS DNA samples (N=320) as well as 225 Tasmanian population control samples. Sample reactions were 8 µL with the following per sample:

4 µL	Bioline SensiFAST™ SYBR® No-ROX master mix
0.1 µL	40× TaqMan custom genotyping probe
2.9 µL	H_2O
1 µL	Sample DNA at 10 ng/µL

Thermal cycling was conducted on a LightCycler® 480 system in a 96 well plate format with the following cycling conditions:

95ºC 10 mins
45 cycles of:
95ºC 10 secs
60ºC 30 secs
Signal acquisition
72ºC 1 sec
40ºC hold

Analysis was conducted in the LightCycler® 480 software, version 1.5.0 to determine sample genotypes. Confirmation by Sanger sequencing was used for samples identified as variant carriers, and those yielding ambiguous genotyping results.

4.3.1.7 Statistical analysis of TaqMan genotyping results

To calculate whether there is a statistical enrichment of the identified variants in the TFHMS resource allelic odds ratios were calculated, using a two-tailed Fisher's exact test where sample size permitted. Alternatively a two-tailed Chi-square test with Yates' correction was used. The variant allele frequencies in the TFHMS samples

were compared to the frequencies in Tasmanian population controls, and also separately to the Exome Variant Server (EVS) European American population samples[271]. The null hypothesis for each analysis was that the occurrence of the minor allele of the variant in the TFHMS samples was not enriched in comparison to the control samples. As the TFHMS samples contained related individuals in different families each family was only counted once, regardless of the number of people genotyped in that family. When the family contained a carrier of the minor allele this individual was chosen, otherwise the chosen individual was homozygous for the major allele. Exceptions were made in large families with multiple carriers of the minor allele that were separated by greater than eight meioses. This reduced the TFHMS samples from N=320 to N=137. Statistical tests were calculated using Prism 6 for Mac OS X, version 6.0f from GraphPad Software, Inc.

4.3.1.8 Illumina MiSeq Nextera® XT DNA custom amplicon sequencing of TNFSF9

Overlapping amplicons, 1 kb to 5 kb in length, were designed to encompass exons, introns and regulatory regions of *TNFSF9* using NCBI's Primer-BLAST[270]. For primer design, known SNPs and repetitive or low complexity regions were taken into account and avoided by Primer-BLAST. A minimum PCR product size of 1 kb and a maximum of 5 kb were selected with specificity tested against the human genome. The primer sequences, optimised PCR reaction conditions and thermal cycling conditions used for each amplicon are detailed in Appendix 4.2.

DNA samples from 96 TFHMS cases were selected for amplicon sequencing of *TNFSF9*. These samples were selected on the basis of disease subtype and age at diagnosis with preference firstly for B-cell and T-cell HMs, to match the subtypes of HMs diagnosed in the *TNFSF9* mutation carriers, and then a preference for younger ages of diagnosis. Sample details are located in Appendix 4.3. For each sample, 25 ng of DNA was amplified for each PCR amplicon. Each PCR product was assessed by agarose gel electrophoresis and samples where amplicons failed to amplify were re-amplified where possible. For each amplicon six samples were quantitated using a Qubit 2.0 fluorometer system with the high sensitivity (HS) assay, according to the manufacturer's instructions. The measured concentrations were averaged per amplicon to represent the approximate PCR product concentration of each PCR

amplicon plate of 96 samples. For the Nextera® XT sequencing, approximately equal amounts of each amplicon were pooled together per sample. 12 pooled amplicon samples were quantitated, again using the Qubit HS assay, and the concentration was averaged across all samples to obtain an approximate global pooled amplicon concentration. The global pooled amplicon concentration was then serially diluted to 0.2 ng/μL. This dilution was then quantitated using the Qubit HS assay to confirm that all samples were approximately 0.2 ng/μL. 5 μL (1 ng) of each pooled amplicon sample was used in the Nextera® XT kit following the manufacturer's October 2012 protocol. Briefly pooled amplicon samples were fragmented, into approximately 300 bp fragments, and tagged with adapter sequences using transposomes. Index sequences per sample were applied in a reduced-cycle PCR amplification creating fragments to facilitate pooling of all 96 samples ready for high-throughput sequencing on the MiSeq.

Paired end 300 bp read length sequencing was conducted on an Illumina MiSeq system. The MiSeq software automatically separated samples according to the sample specific indexes and generated a FASTQ data file for each sample. FASTQ files were aligned to the hg19 reference genome and variants called as described in section 3.3.2. Sequencing coverage analysis was performed for the target region as described in section 3.3.3. Variants were filtered and annotated using ANNOVAR as described in section 3.3.5.

4.4 Results

4.4.1 Sanger sequencing confirmation of nineteen variants

The nineteen prioritised variants, resulting from the tiered analysis in Chapter 3, from families LK0051, LK0124, LK0139 and LK0153, underwent laboratory confirmation by Sanger sequencing. This confirmed thirteen of the variants as true positives from the NGS analysis, as shown in Table 4.1. The variants in *NOTCH1*, *PABPC1* and *SPHK2* could not be confirmed. Additionally the *PDE4DIP* variant was unable to be confirmed by Sanger sequencing as specific primers could not be designed to this region. Accordingly a custom TaqMan genotyping probe was designed but this probe was unable to differentiate between variant carriers and non-carriers.

With thirteen variants in three families verified by Sanger sequencing, together with time and funding constraints, a secondary subjective prioritisation was needed to select two variants for more extensive laboratory analysis. Variants from LK0051 were focused on first because of the clear inheritance pattern between the uncle-nephew pair and their unaffected female relative; this narrowed selection to five variants in five genes. From these, based on the potential relevance of gene function to HM development; the novelty of the association between these genes and HM predisposition; and the specific interesting amino acid changes; the variants in *TDP2* and *TNFSF9* were selected for high-throughput genotyping using custom designed TaqMan genotyping probes.

Table 4.1 Sanger sequencing confirmation of selected variants.

Family	Gene	Variant / dbSNP138 ID if known	Sanger sequencing confirmed
LK0051	*LRP5*	chr11:68181292 C>A	Yes
LK0051	*MMP8*	chr11:102585288 T>G rs138686754	Yes
LK0051	*PEX6*	chr6:42934551 G>A	Yes
LK0051	*TDP2*	chr6:24658126 C>T rs200729372	Yes
LK0051	*TNFSF9*	chr19:6534728 G>C rs61750000	Yes
LK0051, LK0124, LK0139, LK2042	*NOTCH1*	chr9:139417464 T>G	No
LK0051, LK0124, LK2042	*NF1*	chr17:29508805 T>G rs200962248	No
LK0124	*STT3B*	chr3:31659458 T>G rs199778452	No
LK0139	*DUSP10*	chr1:221874839 TG>T	Yes
LK0139	*PDE4DIP*	chr1:144863438 G>T rs140993521	No, unable to design specific primers
LK0139	*SPHK2*	chr19:49132198 A>C rs200347384	No
LK0139, LK2042	*PABPC1*	chr8:101721839 C>A rs202074479	No
LK0153	*GIT1*	chr17:27904190 G>A rs202085570	Yes
LK0153	*HAL*	chr12:96371767 A>G rs150591434	Yes
LK0153	*MET*	chr7:116381047 A>G rs374733251	Yes
LK0153	*NAT10*	chr11:34152973 C>A rs146685334	Yes
LK0153	*NID2*	chr14:52509033 G>A rs143412278	Yes
LK0153	*RARS*	chr5:167929060 CCTT>C	Yes
LK0153	*RIPK2*	chr8:90796371 GTAA>G	Yes

4.4.2 TaqMan genotyping of TDP2 rs200729372 and TNFSF9 rs61750000

The *TDP2* rs200729372 and *TNFSF9* rs61750000 variants were examined using custom TaqMan genotyping probes. The variants were validated as heterozygous using this method in the known carriers LK0051-001, LK0051-007 and LK0051-128, as was confirmed by Sanger sequencing. Figures 1 and 2 in Appendix 4.4 show representative TaqMan genotyping plots with two of the three known variant carriers (LK0051-001 and LK0051-007) genotyped on the assay plate shown along with 92 other samples homozygous for the reference allele.

For *TDP2* rs200729372, using TaqMan genotyping, one additional heterozygous variant carrier was identified as listed in Table 4.2 and was an unaffected daughter of a familial case with NHL. Sanger sequencing confirmed the identified variant carrier.

Table 4.2 *TDP2* rs200729372 additional variant carrier identified through TaqMan genotyping screen of TFHMS samples and population controls.

ID	Details	Confirmed by Sanger sequencing
LK0124-209	Unaffected daughter of a familial HM case with NHL (DNA unavailable)	Yes

For *TNFSF9* rs61750000, a further eight additional heterozygous variant carriers were identified as listed in Table 4.3. One of these was a population control sample (out of 225 population controls screened), resulting in a minor allele frequency (MAF) of 0.002 in this small population. This is lower than the 0.005 MAF from the EVS European American population data for this variant[271]. The remaining seven variant carriers were HM cases or unaffected relatives from the TFHMS sample set, as described in Table 4.3. Sanger sequencing confirmed the identified variant carriers.

Table 4.3 *TNFSF9* rs61750000 additional variant carriers identified through TaqMan genotyping screen of TFHMS samples and population controls.

ID	Details	Confirmed by Sanger sequencing
DVA1746	Population control, no genealogical evidence of a family history of disease	Yes
LK0001-094	Unaffected offspring of a MM patient (DNA unavailable)	Yes
LK0016-004*	Unaffected sibling of five familial CLL cases (DNA unavailable)	Yes
LK0016-005*	Unaffected offspring of familial HM case with CLL (DNA unavailable)	Yes
LK0016-007*	Unaffected offspring of familial HM case with CLL (DNA unavailable)	Yes
LK0016-008*	Unaffected offspring of familial HM case with CLL (DNA unavailable)	Yes
LK0016-187*	Familial HM case with NHL nodular	Yes
LK0628-001	Non-familial DLBCL case	Yes

*LK0016-005, LK0016-007, LK0016-008 are siblings, and LK0016-004 is their uncle and sibling of their CLL affected parent. LK0016-187 is a distant relative, more than eight meioses away.

4.4.3 Statistical analysis of population screening of *TDP2* rs200729372 and *TNFSF9* rs61750000

To determine whether carriers of the variants identified in *TDP2* and *TNFSF9* from analysis of the LK0051 family are statistically enriched in the TFHMS resource, in comparison to control populations, odds ratio calculations were performed. Control populations for comparison were the 225 Tasmanian population controls in one analysis, and a second analysis conducted using a population control data set drawn from the EVS European and American populations[271]. As shown in Table 4.4, carriers of the *TDP2* rs200729372 variant are enriched in the TFHMS resource when compared with the EVS samples (P=0.0028, OR=63.14 [5.71−698.80]), but not the Tasmanian population control samples (P=0.14), using a Fisher's exact test.

For rs61750000 in *TNFSF9*, as shown in Table 4.5, carriers of this variant were enriched in the TFHMS resource, when compared to both the EVS samples (P=0.01, OR=3.78 [1.49−9.64]), and the Tasmanian population control samples (P=0.03,

OR=8.35 [0.96−71.85]). Due to the larger number of variant carriers in the EVS samples a Chi-square test with Yates' correction was performed, whereas the Fisher's exact test was used for the comparison with the Tasmanian population control samples.

Table 4.4 Statistical analysis of population screening of *TDP2* rs200729372.

	Minor allele (T)	Major allele (C)	P-value	OR [95% CI]	Test
TFHMS	3	271			
Tasmanian population controls	2	448	0.14	8.27 [0.39−172.90]	Fisher's exact test
EVS European / American population controls	1	8587	0.0028*	63.14 [5.71−698.80]	Fisher's exact test

*Indicates a statistically significant finding

Table 4.5 Statistical analysis of population screening of *TNFSF9* rs61750000.

	Minor allele (C)	Major allele (G)	P-value	OR [95% CI]	Test
TFHMS	5	269			
Tasmanian population controls	1	449	0.03*	8.35 [0.96−71.85]	Fisher's exact test
EVS European / American population controls	42	8548	0.01*	3.78 [1.49−9.64]	Chi-square with Yates' correction

*Indicates a statistically significant finding

4.4.4 Custom amplicon sequencing of *TNFSF9*

With evidence that the rs61750000 variant in *TNFSF9* was enriched in the TFHMS, when compared to both Tasmanian population controls and the EVS population controls, *TNFSF9* was selected for whole gene screening in 96 HM cases from the TFHMS. Overlapping PCR amplicons were designed to cover the *TNFSF9* gene. These amplicons were amplified in 96 HM cases from the TFHMS, as described in Appendix 4.3, including LK0051-001 as a known positive control, to screen for

variants in *TNFSF9* other than rs61750000 that may be contributing to HM predisposition. HM cases were selected on the basis of being a familial case, or non-familial cases with similar subtypes to the known variant carriers, LK0051-001 and LK0051-128, a T-cell non-Hodgkin lymphoma and a B-cell Burkitt lymphoma respectively. Figure 4.1 shows the sequencing depth coverage of the *TNFSF9* in each sample, calculated as per section 3.3.3. Two samples, LK0004-012 and LK0024-001, had low coverage with less than 50% of the gene covered at 10× depth or greater. Excluding these two samples, the other 94 samples have coverage of greater than 10× depth for over 80% of the target region.

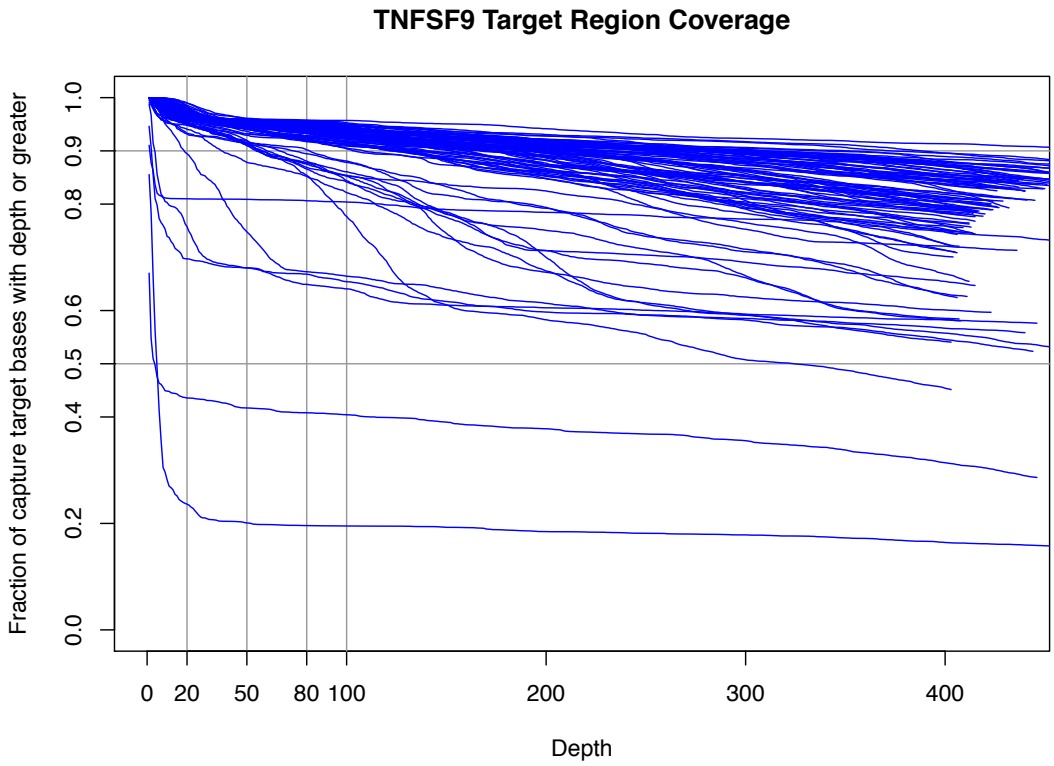


Figure 4.1 *TNFSF9* coverage plot.
Calculated in BEDTools and plotted in R this plot shows the fraction of bases (Y-axis) at a coverage depth (X-axis). Two samples have low coverage while the remaining 94 samples have ≥ 80% of the region with ≥ 10× read depth.

Analysis of the custom amplicon sequencing data identified the known *TNFSF9* rs61750000 variant in LK0051-001 who was sequenced as a positive control as well as LK0016-187 and LK0628-001 as previously identified by TaqMan genotyping (Table 4.3). After variant filtering as per section 3.4.2 there were nineteen variants identified across the 96 samples, as shown in Table 4.6. As per section 3.3.7.3, when the CADD scoring model is applied to these variants, none are predicted to be deleterious using a CADD phred-like scaled C score threshold of ≥ 10. All variants are non-coding variants in the upstream and downstream regions of *TNFSF9*. No additional exonic variants, or variants predicted to be deleterious using CADD, were identified.

Table 4.6 Summary of additional *TNFSF9* variants identified through gene sequencing.

Variant / dbSNP 138 ID if known	Gene location	CADD phred-like scaled C score	Number of variant carriers
chr19:6530040 GACAT>G	Upstream	7.175	40
chr19:6530956 G>A rs2234171	Upstream	6.721	1
chr19:6530033 ACAG>A	Upstream	6.229	2
chr19:6530103 T>C	Upstream	6.188	2
chr19:6530037 ACAG>A	Upstream	6.168	40
chr19:6530038 CAGA>C	Upstream	6.168	40
chr19:6535291 ATATATAT>A	3'UTR	6.166	2
chr19:6530020 CAGAG>C	Upstream	6.106	1
chr19:6535078 C>T rs75982228	3'UTR	5.023	1
chr19:6530036 GACA>G	Upstream	4.74	40
chr19:6535801 C>T rs3865470	3'UTR	2.87	1
chr19:6530034 C>CACA	Upstream	2.479	6
chr19:6530035 AGAC>A	Upstream	1.45	40
chr19:6536461 G>A	Downstream	1.016	1
chr19:6535297 AT>A	3'UTR	0.865	2
chr19:6530027 AAAG>A	Upstream	0.814	3
chr19:6530032 G>GACAC	Upstream	0.761	7
chr19:6530024 GAGAA>G	Upstream	0.755	5

4.5 Discussion

4.5.1 Validation of sequencing variants

Sanger sequencing was used for validation of the nineteen prioritised variants that resulted from the family-based analyses of the LK0051, LK0124, LK0139 and LK0153 families in Chapter 3. Thirteen of the nineteen variants were validated. It is interesting to note that of the five variants that failed validation, three were identified across multiple families. The remaining variant of the nineteen in *PDE4DIP* was unable to be validated by Sanger sequencing or by using a custom TaqMan genotyping probe. This variant is located in a region of high homology to other genomic locations, explaining the difficulty in designing specific PCR primers. Further optimisation is required to genotype this variant. The remaining variants that failed validation in *NF1*, *NOTCH1*, *PABPC1*, *SPHK2* and *STT3B* are likely to be artefacts of the NGS process, or result from errors in sequence alignment and variant calling. The identification of these false positives highlights the need for variant validation by an alternate method such as Sanger sequencing.

The occurrence of false positives is an expectation in NGS data as the current technologies have error rates ranging from one error in one thousand nucleotides (99.9%) to one error in ten million nucleotides (99.9999%) as reviewed in Robasky *et al.*[269]. With limitless funding false positive rates could, hypothetically, be reduced by sequencing multiple replicates of the same sample across multiple NGS technologies. However as such an approach is not typically feasible, read depth and base call quality are relied upon, as was used in this study. Setting a higher threshold for these values may reduce the number of false positives, at the risk of removing true variants. Sharing analyses, as was used to identify the prioritised variants in this chapter, also assists in reducing the number of false positives, as identification of the same variant across different samples increases support for this variant being a true positive.

4.5.2 Resource genotyping of selected prioritised and validated variants

With thirteen variants validated by Sanger sequencing from NGS analysis and prioritisation in three of the five TFHMS families analysed, the two variants with the strongest biological evidence supporting a role in HM susceptibility in family LK0051 were selected for genotyping across the TFHMS resource (N=320) and a set

of Tasmanian population controls drawn from the case-control study of prostate cancer outlined in section 2.2.3 (N=225). The variants selected were the rs200729372 C>T variant in *TDP2*, based on the DNA damage repair role of the encoded protein[272], and the rs61750000 G>C variant in *TNFSF9* based on the established role of the encoded protein in B-cell and T-cell biology[273].

4.5.2.1 TDP2 rs200729372 C>T

Genotype screening of the *TDP2* rs200729372 variant in the TFHMS resource and Tasmanian population controls identified this variant in the unaffected daughter of a familial HM case with NHL (DNA of the case was unavailable) in addition to the LK0051 family. Statistical analysis of the allele frequencies of this variant using a Fisher's exact test did not identify enrichment of the variant in the TFHMS resource in comparison to the genotyped Tasmanian population controls. However in comparison to the allele frequency in the EVS Caucasian population (N=4295) a P-value of 0.0028 and an OR of 63.14 [5.71−698.80] was found. This finding provides supportive evidence that the variant may be contributing to HM predisposition. Even so, a larger number of Tasmanian population controls should be screened for this variant. As well as genotyping of other HM cases, for this and other *TDP2* variants, particularly B-cell and T-cell HM subtypes given the occurrence of this variant in TFHMS cases with these types of HMs. The connection between this *TDP2* variant and HM development is a novel finding. Additionally, it is interesting to note that this variant appears only in Caucasian populations in large public databases of genetic variation, including the EVS[271], 1000GP[236] and the recent Exome Aggregation Consortium (ExAC)[274] populations. This is consistent with the higher prevalence of HMs in Caucasian populations, as discussed in Section 1.4.

The *TDP2* variant causes an amino acid change converting a serine residue to an asparagine residue at position 144 of the tyrosyl-DNA phosphodiesterase 2 protein. The asparagine residue is a larger amino acid with different hydrophobicity properties to the serine residue. While the specific protein structure of human TDP2 is unknown, the homologous mouse Tdp2 protein structure has been determined[275] which allowed structural modelling of this variant using the Project HOPE interface[250]. Located on the surface of the protein the S144N amino acid change may disturb hydrophobic

interactions with other molecules at the protein surface, however molecular contacts with this specific residue have not been identified. The change occurs within the phosphodiesterase protein family domain of TDP2. This domain is responsible for the repair of topoisomerase-induced DNA double strand breaks by cleaving the covalently bound topoisomerase at the 5' end of the DNA break, freeing the end for DNA repair[272].

The DNA repair role of TDP2 has been shown to occur through error-free non-homologous end-joining[276]. Transgenic mice with a *Tdp2* deficiency do not show an overt phenotype[276]. However, when challenged with the topoisomerase inhibitor and anti-cancer drug etoposide, mice with a *Tdp2* deficiency have an increase in genome instability in haematopoietic cells in the bone marrow[276]. This is evidence that TDP2 has a role in maintaining genomic stability through its interactions with topoisomerase. It is possible that the S144N mutation in the LK0051 family causes a dysfunctional TDP2, which increases genome instability and contributes to HM predisposition. Furthermore, publically accessible microarray gene expression profiling data from the GeneAtlas dataset[252,253] for *TDP2* shows that this gene has elevated mRNA expression in haematopoietic cells, as well as the small intestine and colon (Appendix 4.5 A). Increased expression in these tissues suggests a biological role for this gene. The elevated expression in the small intestine and colon is interesting in light of the HM subtype of one of the carriers of this variant, LK0051-128, who has BL of the small intestine. This suggests that TDP2 activity may be important in haematopoietic tissues, particularly those of the small intestine, and deregulation may have contributed to a HM predisposition in variant carriers.

4.5.2.2 TNFSF9 rs61750000 G>C

Genotype screening of the *TNFSF9* rs61750000 variant in the TFHMS resource and Tasmanian population controls identified this variant in a non-familial DLBCL case, one other familial HM case with NHL and five unaffected first-degree relatives of familial cases. Statistical analysis of the allele frequencies of this variant identified evidence for enrichment of the variant in the TFHMS resource in comparison to both the genotyped Tasmanian (N=225) and the EVS population controls (N=4295) with P-values of 0.03 (OR=8.35 [0.96−71.85]) and 0.01 (OR=3.78 [1.49−9.64])

respectively. Additionally, it is interesting to note that this variant appears in Caucasian, African and Latino populations, but not Asian, in large public databases of genetic variation, including the EVS[271], 1000GP[236] and the recent Exome Aggregation Consortium (ExAC)[274] populations. This is consistent with the higher prevalence of HMs in Caucasian populations, as discussed in Section 1.4 and also raises the question of whether this variant may also contribute to specific HMs in African or Latino populations, particularly given that BL, related to EBV infection, is endemic to Africa[63].

With statistical evidence implicating the *TNFSF9* variant rs61750000 in HM development a set of 96 HM cases from the TFHMS were screened for additional *TNFSF9* variants that may be contributing to HM development, using custom amplicon sequencing of the gene. Several variants were detected, however none were predicted to have an effect on gene function, leaving rs6175000 as the most deleterious variant identified in *TNFSF9*. It is possible that screening more HM cases, specifically cases with Burkitt lymphoma or T-cell non-Hodgkin lymphoma, would identify additional variant carriers of the rs61750000 variants or other coding or regulatory variants that impact upon gene function.

The rs61750000 variant results in a glycine to alanine amino acid change at residue 139 (G139A) of the TNF ligand superfamily member 9 protein, within the TNF protein domain. The TNFSF9 protein occurs as a homotrimer of three TNFSF9 subunits in a 'propeller' conformation[277]. The glycine to alanine residue change in terms of amino acid properties is a very conservative change; the alanine residue differs from the glycine by a single methyl group. However, the change occurs at a residue that is completely conserved across all homologous alignments of the TNF protein domain (Appendix 4.6 Figure 1), indicating that the glycine amino acid at this location is biologically important and a small change may have a significant impact on protein function. The change also occurs nearby two residues that form part of the trimer interface between the TNFSF9 subunits, potentially impacting on formation of the TNFSF9 homotrimer (Appendix 4.6 Figure 2). Indeed the introduction of a V140A mutation directly next to the G139A prevents homotrimer formation *in vitro*[277].

TNFSF9 is an extracellular transmembrane protein[277]. It is the ligand for the TNFRSF9 receptor, also expressed on the cell membrane and is able to act as a bidirectional signalling molecule[273]. Functionally, TNFSF9 is mainly expressed on antigen presenting cells including dendritic cells, monocytes, macrophages and B-cells[273]. It is able to co-stimulate proliferation of T-cells through the TNFRSF9 receptor and can signal back into the antigen presenting cells inducing proliferation, prolonging cell survival and enhancing the secretion of proinflammatory cytokines[273]. However, when expressed on T-cells, TNFSF9 when stimulated leads to an inhibition of T-cell proliferation and promotes apoptosis[278,279]. This highlights the ability of TNFSF9 to affect opposite signalling pathways depending upon the cell type involved. Evidence implicating the involvement of *TNFSF9* in HM development includes the finding of recurrent deletion of the gene in DLBCL and BL[280] and the development of B-cell lymphomas in 60% of transgenic *Tnfsf9* deletion mice by one year of age[281]. Furthermore gene expression profiling data from the GeneAtlas dataset[252,253] for *TNFSF9* shows elevated expression in B-cells and two HM cell lines; the BL DAUDI cell line and the APML HL60 cell line (Appendix 4.5 B). How the observed increase in *TNFSF9* expression in the HM cell lines reconciles with the loss of expression in *TNFSF9* transgenic knock-out mice and human HMs remains to be determined. However, it is possible that, paradoxically, both loss of function and gain of function in *TNFSF9* contributes to HM development, as is known for *RUNX1*[282]. Taken together, these data provide compelling evidence that supports a potential role for deleterious variants in *TNFSF9* contributing to the biology of familial HM predisposition.

The connection between a germline mutation in *TNFSF9* and HM predisposition is a novel finding. The plausible biological links that support the role of *TNFSF9* in the development of familial HMs present a new avenue of research in this area. It is possible that the *TNFSF9* rs61750000 variant contributed to HM development in LK0051-001 and LK0051-128 in combination with the other candidate variants identified. Functional validation of the effect of this variant on TNFSF9 protein activity, together with identification of this, or other, recurring deleterious *TNFSF9* variants in familial and non-familial HM cases will provide the supporting evidence to implicate disruption of this gene as a causal factor in HM development.

4.6 Conclusion

This chapter described the validation experiments of nineteen prioritised variants implicated in HM development from NGS analysis of four TFHMS families. Thirteen variants were validated by Sanger sequencing as true variants. Two of these variants, rs200729372 in *TDP2* and rs61750000 in *TNFSF9*, from the LK0051 family, because of compelling biological evidence implicating the potential involvement of these genes in HM development, were genotyped in the full 320 individuals of the TFHMS resource and 225 Tasmanian population controls. Statistical analyses of allele frequencies identified the *TDP2* variant as statistically enriched in the *TFHMS* resource in comparison to EVS Caucasian controls, but not Tasmanian population controls, whereas the *TNFSF9* variant was enriched in TFHMS in comparison to both control groups. Screening of the *TNFSF9* gene by amplicon-based NGS in 96 TFHMS cases did not identify any other deleterious variants in addition to rs6175000.

This is the first report of a connection between HM susceptibility and germline variation in *TDP2* and *TNFSF9*. The analyses conducted here are the first stage in establishing a causal role for variants in these genes and HM predisposition. These novel findings suggest that further studies of the biological effect of these specific variants, and these genes, in the context of HM development are warranted.

Chapter 5 - Variance components modelling of telomere length in TFHMS

5.1 Preface

Sections of this chapter were published in *Oncology Reports* in 2015. The article, entitled 'A retrospective examination of mean relative telomere length in the Tasmanian Familial Haematological Malignancies Study'[1], is included in this dissertation as Appendix 5.1.

5.2 Introduction

Telomeres are hexameric nucleotide repeat sequences of TTAGGG found at the ends of chromosomes[283]. The primary role of the telomere is to cap chromosome ends to prevent aberrant recombination as a result of exposed chromosomal DNA making telomeres essential for maintenance of genomic integrity[284]. Due to the incomplete DNA replication of chromosome ends by DNA polymerases, with each cell division telomeres shorten, eventually triggering cell senescence or apoptosis to prevent further shortening and exposure of chromosomal DNA[285]. While the telomerase complex counteracts telomere shortening in actively dividing cells by catalysing the addition of TTAGGG repeats to chromosome ends[285,286], telomeres nevertheless progressively shorten with age[285,287,288].

5.2.1 Telomere length in haematological malignancies

There is accumulating evidence that implicates telomere length as an important factor in the development of a range of HMs. As discussed in section 1.5.4.1 inherited changes in telomere length as a result of mutations in the telomere complex genes *TERT* and *TERC* have been identified in four HM families with MDS-AML[147]. Other studies have identified shorter telomeres in circulating tumour cells of patients with AML[289], cutaneous T-cell lymphoma[290], MM[291], MPNs[292], APML[293] and ALL[294]. While these studies have revealed important insights into telomere dynamics in circulating tumour cells there has been less focus on pre-disease and remission

telomere lengths of non-malignant blood cells in HM patients. Indeed, one prospective study of telomere length in pre-disease blood samples surprisingly showed that longer telomere length was associated with a future risk of NHL[295]. Further, in a study of chronic leukaemia, Mansouri and colleagues elegantly show that telomere length has potential as a clinical prognostic marker in HMs[296]. In their study patients with shorter telomeres were associated with high-risk genetic markers and in patients with otherwise good prognostic markers, telomere length was an independent prognostic factor that subdivided the good prognosis group into groups with distinct outcomes. Therefore there is potential for telomere length to be a clinically relevant prognostic risk factor for HMs.

The compounding effects of disease and chemotherapy regimens may complicate retrospective studies as it has been suggested that these factors contribute to telomere length shortening. Although, examination of telomere length in APML showed that whilst telomere length changed over the course of disease and chemotherapy, when complete remission was achieved there was no difference between telomere length in patients and controls[293]. A recent larger study in a chronic leukaemia, CLL, that also sequentially followed patients over disease course and treatment showed that telomere length at diagnosis remained unaffected by chemotherapy in remission[296].

5.2.2 Telomere length as a quantitative trait in cancer

A number of studies have reported an association between telomere length, as a quantitative trait in lymphocytes, and an increased risk of age-related diseases, including cancer[297,298]. Excessive telomere shortening leads to increased genetic instability, decreased chromosomal stability and chromosome end-to-end fusions[284,299], which can lead to a malignant cell transformation[299]. Thus telomere length is a proposed risk factor for cancer given its importance in maintaining genomic integrity. Shorter telomeres have been associated with a range of cancers including cancers of the head and neck, bladder, lung and renal cells[297,298,300].

Studies of monozygotic and dizygotic twins and large families have indicated a genetic component to the determination of telomere length. Estimates of the heritability of telomere length as a trait range between 78% and 82% in studies of

twins and sibling pairs[301,302] and 44% in a study of large Amish families[303]. Although it has been proposed that the heritability of telomere length can be accounted for by shared environmental factors[304] the consensus is that telomere length is primarily determined by parental inheritance including at least partial inheritance of chromosome specific telomere lengths[305,306]. This view is strongly supported by mouse models of telomere length inheritance[307].

Telomere length, as a quantitative trait, can be measured by a range of different techniques including flow-FISH, single telomere length analysis (STELA), southern blotting and quantitative PCR-based methods[308]. For many years southern blotting was viewed as the gold standard for measuring telomere length[309]. The development of the monochrome multiplex quantitative PCR-based assay has introduced a method of measuring telomere length that has many practical and analytical improvements on southern blotting, and in addition, shows linear correlations with the original technique[310,311]. This assay has facilitated large-scale quantitative studies of telomere length in a variety of contexts.

5.2.3 Quantitative traits

Quantitative traits, such as telomere length, are useful as disease markers. As explained by Almasy[312], identifying the sources of variance in a quantitative trait between individuals, identifies why they are different to each other. The two sources of variance are genotype and environment, that is, variance in a trait is the result of a genetic contribution and/or an environmental contribution (or their interaction). The heritability of a trait is a measure of the genetic contribution to the variation in the trait. A trait that is highly heritable will therefore be driven primarily by genetics, whereas low heritability can suggest that environmental factors such as lifestyle are a larger contributor to variation in the trait.

One method of examining the factors or covariates that contribute to variation in a quantitative trait like telomere length is to measure the trait in large families with known pedigree structures[313]. With the pedigree structure comes relationship information from which kinship coefficients can be calculated. As shown by Almasy, first degree relative pairs (siblings, parent-offspring) will share 50% of their genetics

with the degree of sharing halving with each degree of familial separation[312,313]. People related by multiple lines of descent, through consanguineous relationships for example, will share more than the same relationship without multiple lines of descent. These kinship coefficients, together with the phenotypic information from the quantitative trait, allow the calculation of the trait's heritability. If a trait is heritable it follows that more closely related relatives, due to shared genetics, will have more similar trait values with distant relatives still having a closer trait correlation than two unrelated individuals. In this way family studies have a much greater power for detecting the genetic contributions to trait variation than population studies of unrelated individuals. In a population study the trait of interest may be influenced by many different genes, specific to each person, which may dilute rare and low frequency genetic variants that have a large effect on trait variation to an undetectable level[186]. Whereas when studying a family or series of families, the analysis is based on related individuals, so there is enrichment for the same genetic variants due to identical by descent (IBD) flow of chromosomes through the pedigree. The dilution that is seen in population studies is therefore avoided. This enhances the ability to detect factors contributing to trait variation[186].

Using the method of variance components modelling, accounting for kinship, the degree to which various covariates contribute to the variation in the trait can be assessed[313]. A significantly associated covariate indicates that changes in that covariate account for part of the measured difference in the trait. Covariates can be factors such as age, sex, disease status or genotype data. This means that variance components modelling analysis can be used to identify phenotypic covariates and genetic variants that account for trait variation.

Given the evidence supporting a role for telomere length and genes influencing telomere biology in cancer and HMs, as well as its heritability as a quantitative trait, telomere length was examined in the TFHMS resource. Examining telomere length in HM families may identify factors related to HM disease development and inform biomarkers for disease risk estimation. Variance components modelling was employed to analyse telomere length in the TFHMS resource, to identify new evidence of whether telomere length is involved in HMs.

5.3 Aims

The aim of this study was to explore in the TFHMS resource whether telomere length as a quantitative trait is involved in familial HMs and to find new evidence supporting telomere length as a prognostic risk factor for HMs.

5.4 Methods

5.4.1 Study samples for telomere length measurement

Telomere length was measured in DNA extracted from peripheral blood samples, as per section 2.2.4, from 55 familial HM cases, 191 unaffected relatives of familial cases and 75 non-familial cases. DNA from 40 TFHMS families was available for this study with samples available from both HM cases and unaffected relatives in 14 families. The remaining families were comprised of samples from HM cases with a known family history of disease alone or from unaffected relatives of HM cases. Of the 191 unaffected relatives, 171 were first-degree relatives of HM cases and the remainder were more distantly related or spouses. For HM cases, DNA was collected from 1 month to 64.9 years post HM diagnosis (mean 9.9 years). Population controls (N=758) are detailed in section 2.2.3.

5.4.2 Telomere length measurement

Mean relative telomere length was measured in peripheral blood samples using a slightly amended protocol for a validated monochrome multiplex quantitative PCR method outlined by Cawthon[311]. This method measures the relative telomere length by calculating the ratio, T/S, between telomere repeat copy number amplification (T) and the amplification of a single-copy gene, albumin *ALB* (S). The average T/S ratio was obtained as the mean of the triplicate measurements for each sample. Individual measurements were excluded from average T/S ratio calculation when the replicate failed or a large standard error was observed. The coefficient of variation calculated across all assay plates using repeated cross-plate samples was 3.4%.

Telomere length measurement was performed in 10 µl reaction volumes using a LightCycler480 in a 96 well plate format. Each 96 well plate contained a six point standard curve 2 ng, 5 ng, 15 ng, 50 ng, 100 ng and 150 ng, a unique sample common

to each plate, a no template control and 24 unknown case / controls samples all repeated in triplicate, with 1.6% sample replication across plates. The genomic DNA used for the standard curve was from a 27-year-old female control study participant.

Final reagent concentrations per sample were:

5 ng	Genomic DNA,
200 nM	primer telg (5'-acactaaggtttgggtttgggtttgggtttgggttagtgt-3')
700 nM	primer telc (5'-tgttaggtatccctatccctatccctatccctatccctaaca-3')
500 nM	primer albu (5'-cggcggcgggcggcgcgggctgggcggaaatgctgcacagaatccttg-3')
500 nM	primer albd (5'-gcccggcccgccgcgcccgtcccgccggaaaagcatggtcgcctgtt-3')
0.625 U	Amplitaq gold (Applied Biosystems Inc.)
GeneAmp 10×	PCR buffer (Life Technologies) containing:
	50 mM KCl
	10 mM Tris-HCl pH 8.3
	1.5 mM $MgCl_2$
1 mM	DTT
1 M	Betaine (Sigma-Aldrich)
0.0025 mM	Syto 9 (Life Technologies)
0.25 mM	Of each dNTP (Bioline)

Thermal cycling conditions were as follows:

95°C 15 mins
Two cycles of:
94°C 15 secs
49°C 60 secs
Four cycles of:
84°C 20 secs
59°C 30 secs
40 cycles of:
94°C 15 secs
59°C 30 secs
Signal acquisition for telomere repeat copy number amplification
84°C 30 secs
85°C 20 secs
Signal acquisition for albumin amplification

A melt curve was generated for each plate. CT values were calculated using LinRegPCR[314] and a standard curve was generated for both the telomere and albumin PCRs. A linear regression of the standard curve measurement values was used to correct for any variation in fluorescence levels derived from small fluctuations in DNA concentration. The equations from the linear regression of each standard curve

were then used to calculate the log(DNA) value for the unknown case / control samples.

5.4.3 Statistical analysis

Average T/S ratios greater than 4 standard deviations from the control mean were excluded as outliers. For analysis, to prevent non-normal distribution errors, mean relative T/S ratios were transformed to fit a normal distribution using the inverse-normalisation option in SOLAR (version 6.6.2)[312,313]. In order to fully utilise the extended pedigree study design, correct for relatedness, and to maximise the information provided by telomere length as a quantitative trait we used variance components modelling in SOLAR[312,313] to determine the heritability of telomere length (adjusting for kinship and significant covariates) and to calculate the association between telomere length and disease. The primary benefit to using SOLAR is its ability to incorporate relatedness through the use of a kinship matrix and to fully utilise the quantitative trait data, which increases the power and accuracy of the trait heritability calculation.

Sex, age, age^2 and their interactions were included as covariates in all relevant analyses. Potential batch effects were adjusted for by applying household modelling[312,313], coding each assay plate as a separate household. SOLAR has been used previously in the analysis of telomere length in related individuals[302,303,315]. The algorithms utilised in SOLAR for the analysis of quantitative traits in related individuals are more appropriate to employ in this study than a more traditional approach of analysing quantitative traits using percentiles or quartiles. Nevertheless, observations from a quartile analysis of inverse normalised relative T/S ratios adjusted for age, sex and batch effects in SOLAR with quartiles defined from the adjusted T/S ratios in the control population, are also presented. Bean plots in Figure 1 were constructed using the R package 'beanplot'[316].

5.4.4 Analysis of genetic variants in telomere biology related genes

Variants identified through NGS of 31 TFHMS individuals as described in Chapter 3, section 3.3.5, were used as a variant set to identify variants contributing to shorter telomere length in measured individuals, particularly sequenced HM cases who were

in the lowest quartile of telomere length. The 96,220 variants from section 3.3.5 were filtered as per section 3.3.7.3 to contain variants with CADD phred-like scaled C scores ≥ 10. These are variants within the top 10% of variants predicted by the CADD model to be deleterious[244].

Then, variants in a curated list of genes known to be related to telomere biology were extracted for analysis. The curated list of 162 genes (Appendix 5.2) is built from genes located in the following human pathways from NCBI's Biosystems[317], 'Extension of Telomeres', 'Regulation of Telomerase', 'Telomere Extension By Telomerase', 'Telomere Maintenance', together with 43 telomere biology genes defined by Mirabello *et al.*[318]. The average number of variants of samples in the lowest quartile of telomere length was compared to samples not in this quartile using two-tailed t tests (assuming a Gaussian distribution), conducted in Prism 6 for Mac OS X, version 6.0f from GraphPad Software, Inc.

5.5 Results

5.5.1 Participant characteristics

Families and the numbers of individuals in which telomere length was measured in this study are shown in Table 5.1. Table 5.2 shows the mean age and sex characteristics of the sample groups measured, together with mean relative T/S ratios with 95% confidence intervals. The distribution of HM case subtypes for whom telomere length was measured are summarised in Table 5.3.

Table 5.1 Summary of the TFHMS families used in this study.

Family	Known HM cases	Generations with HM cases	HM cases with telomere length measurement	Unaffected relatives with telomere length measurement
LK0001	14	4	1	16
LK0002	15	3	1	5
LK0004	7	2	1	11
LK0016	18	5	2	19
LK0024	3	2	1	0
LK0026	6	2	1	5
LK0040	7	4	2	2
LK0051	21	5	3	26
LK0054	9	3	0	2
LK0065	8	2	0	8
LK0124	24	5	2	34
LK0132	5	2	0	7
LK0139	7	2	1	2
LK0153	9	2	3	2
LK0511	2	2	1	0
LK0512	2	1	1	0
LK0537	2	1	2	0
LK0546	2	2	1	0
LK0560	2	2	1	0
LK0561	2	2	1	0
LK0600	5	3	2	0
LK0625	4	2	2	0
LK0647	2	2	1	0
LK0672	3	3	1	0
LK0836	6	3	2	5
LK1155	2	1	1	3
LK2042	32	5	6	40
LK2447	3	2	1	2
LK6000	6	2	1	0
LK7739	2	1	1	0
LK7740	2	2	2	0
LK7743	3	2	2	0
LK7744	2	2	0	1
LK7748	2	2	1	0
LK7749	3	2	1	0
LK7750	4	2	2	0
LK7751	9	3	1	0
LK7754	3	1	1	0
LK7755	2	2	1	0
LK7768	2	1	1	0
Non-familial cases	--	--	75	1

Table 5.2 Mean age, sex distribution and relative telomere length in the sample groups.

Sample group	Number	Male sex, No. (%)	Mean age (range)	Mean relative T/S ratio* (95% CI)
Controls	758	578 (76.3)	67.51 (30.67 – 87.97)	0.64 (0.62 – 0.66)
Unaffected relatives of HM cases	191	77 (40.3)	61.65 (27.26 – 92.95)	0.73 (0.69 – 0.76)
All HM cases	130	73 (56.2)	65.14 (13.24 – 95.53)	0.53 (0.50 – 0.56)
Familial HM cases	55	32 (58.2)	64.45 (13.24 – 87.45)	0.57 (0.52 – 0.63)
Non-familial HM cases	75	41 (54.7)	68.79 (22.42 – 95.53)	0.50 (0.46 – 0.53)

*Mean relative T/S ratio is the ratio between telomere repeat copy number (T) and a single-copy gene, *ALB*, copy number (S), a measure of mean relative telomere length

Table 5.3 Disease characteristics of study samples.

HM Subtypes	HM Familial cases No. (%)	HM Non-familial cases No. (%)	All HM cases No. (%)
Acute lymphoblastic leukaemia	2 (3.6)	0	2 (1.5)
Acute myeloid leukaemia	5 (9.1)	8 (10.7)	13 (10.0)
Chronic myeloid leukaemia	0	3 (4.0)	3 (2.3)
Essential Thrombocythemia	1 (1.8)	1 (1.3)	2 (1.5)
Hodgkin Lymphoma	5 (9.1)	4 (5.3)	9 (6.9)
Myelodysplastic syndrome	2 (3.6)	0	2 (1.5)
Myeloproliferative neoplasm	1 (1.8)	2 (2.7)	3 (2.3)
T-cell Non-Hodgkin lymphoma	1 (1.8)	2 (2.7)	3 (2.3)
Mature B Cell Neoplasms			
Non-Hodgkin lymphoma unclassified	2 (3.6)	10 (13.3)	12 (9.2)
Chronic lymphocytic leukaemia	12 (21.8)	12 (16.0)	24 (18.5)
Diffuse large B-cell lymphoma	4 (7.3)	10 (13.3)	14 (10.8)
Follicular Lymphoma	4 (7.3)	9 (12.0)	13 (10.0)
Multiple myeloma	7 (12.7)	5 (6.7)	12 (9.2)
other*	9 (16.4)	9 (12.0)	18 (13.8)
Total	55	75	130

*other includes Burkitt lymphoma, hairy cell leukaemia, lymphoma of mucosa-associated lymphoid tissue and Waldenström macroglobulinemia

5.5.2 Heritability of telomere length in TFHMS

Telomere length in familial and non-familial HM cases, unaffected relatives and control subjects was measured and the heritability of telomere length was calculated to be 62.5% ($P=4.7\times10^{-5}$, SE=0.14). The removal of HM cases (n=130) from analysis only marginally altered the heritability value of mean relative telomere length (75.5%; $P=1.2\times10^{-5}$, SE=0.15).

5.5.3 Variance components modelling analysis of telomere length in TFHMS

The use of variance components modelling in SOLAR permits appropriate statistical analyses inclusive of familial relationships. These analyses revealed that disease status was significantly associated with mean relative telomere length (Table 5.4, primary analysis model 1, $P=2.9\times10^{-6}$) with HM cases having shorter mean relative telomere length when compared with unaffected individuals. We conducted a separate analysis distinguishing familial and non-familial cases. Familial cases and non-familial cases each had significantly shorter mean relative telomere length (Table 5.4; primary analysis model 2, $P=2.2\times10^{-4}$ and 2.2×10^{-5}, respectively).

The most frequent type of HM diagnosed in our study was the mature B-cell neoplasms (MBCNs; Table 5.3). Analysis of MBCNs as one group and HMs other than MBCNs as a second group (combined due to small numbers of other subtypes) showed that both groupings had shorter mean relative telomere length than unaffected individuals from both study families and population controls (Table 5.4; primary analysis model 3, $P=3.5\times10^{-5}$, $P=9.3\times10^{-5}$ respectively). These groupings were then divided according to whether the HM case was familial or non-familial. Analysis showed that all HM case subgroupings maintained significantly shorter mean relative telomere length (Table 5.4; primary analysis model 4). An analysis using specific HM subtypes was not possible due to insufficient statistical power at this level of HM classification with small numbers of HM subtypes (Table 5.3).

Variance components modelling also identified age and sex (Table 5.4; primary analysis model 1, $P=4.8\times10^{-8}$ and 4.0×10^{-3} respectively) as significant covariates for mean relative telomere length variation across all models. Mean relative telomere length declined with age and males had shorter telomeres than females. Age^2 was also

a significant covariate, indicating that the decline of mean relative telomere length with age has a non-linear component (Table 5.3; primary analysis model 1, $P=8.0\times10^{-3}$).

Four sub-analyses of the primary data were also performed to determine whether particular features of the study population were contributing to the disease associations found in the primary analysis models (Table 5.4). Sub-analyses included exclusion of HM cases, controls and unaffected relatives 80 years or older (n=126), exclusion of CLL cases (n=24), exclusion of cases with samples collected within two years of diagnosis (n=24) as well as all three exclusions together (n=162, some individuals were in multiple exclusion categories). In each sub-analysis the principle findings from the primary analysis models were maintained.

Categorisation of cases into quartiles of mean relative telomere length determined from the distribution of age, sex and batch effect adjusted mean relative telomere length in controls (Figure 5.1) shows that 43.1% of HM cases were in the lowest quartile of mean relative telomere length (below lower interquartile dashed line). With 36.4% of familial HM cases and 48% of non-familial HM cases in the lowest quartile, only 13.1% of unaffected relatives were compartively in the lowest quartile. Similarly a low percentage of cases (5.4%) were in the highest quartile of mean relative telomere length (above upper interquartile dashed line) whereas 28.3% of unaffected relatives were in the highest quartile. A clear trend for shorter mean relative telomere length in a higher percentage of HM cases is observed but this analysis does not permit familial relationships to be included in the analysis.

Table 5.4 Variance component modelling analysis of inverse normalised mean relative telomere length - primary analysis and sub-analyses with exclusions.

Models and variables	Primary Analysis P-values	≥ 80 years old excluded (N=126) P-values	CLL cases excluded (N=24) P-values	Possible malignant samples excluded* (N=24) P-values	All exclusions applied (N=162) P-values
Model 1					
age	4.8×10^{-8}	7.5×10^{-5}	1.6×10^{-8}	6.9×10^{-8}	3.4×10^{-5}
age^2	8.0×10^{-3}	0.04	6.0×10^{-3}	0.01	0.07
sex	4.0×10^{-3}	0.01	7.0×10^{-3}	2.0×10^{-3}	0.01
all HM cases	2.9×10^{-6}	7.3×10^{-6}	2.9×10^{-7}	1.1×10^{-4}	4.6×10^{-5}
% trait variance accounted for by model	10.07%	9.46%	10.38%	9.56%	9.38%
Model 2					
age	4.3×10^{-8}	5.1×10^{-5}	1.6×10^{-8}	7.4×10^{-8}	3.7×10^{-5}
age^2	8.0×10^{-3}	0.04	6.0×10^{-3}	0.01	0.07
sex	3.0×10^{-3}	0.01	6.0×10^{-3}	2.0×10^{-3}	9.0×10^{-3}
familial HM cases	2.2×10^{-4}	1.0×10^{-3}	1.6×10^{-5}	0.01	8.0×10^{-3}
non-familial HM cases	2.2×10^{-5}	6.9×10^{-6}	7.1×10^{-5}	3.3×10^{-5}	2.7×10^{-5}
% trait variance accounted for by model	10.62%	10.29%	10.55%	10.48%	10.00%
Model 3					
age	4.7×10^{-8}	7.2×10^{-5}	1.7×10^{-8}	7.7×10^{-8}	3.7×10^{-5}
age^2	8.0×10^{-3}	0.04	6.0×10^{-3}	0.01	0.07
sex	4.0×10^{-3}	0.01	6.0×10^{-3}	2.0×10^{-3}	0.01
MBCNs	3.5×10^{-5}	7.8×10^{-5}	5.5×10^{-6}	5.7×10^{-4}	3.8×10^{-4}
HMs other than MBCNs	9.3×10^{-5}	1.5×10^{-4}	3.4×10^{-5}	1.0×10^{-3}	5.8×10^{-4}
% trait variance accounted for by model	10.08%	9.50%	10.37%	9.56%	9.39%
Model 4					
age	4.8×10^{-8}	4.1×10^{-5}	1.9×10^{-8}	6.6×10^{-8}	3.8×10^{-5}
age^2	9.0×10^{-3}	0.05	7.0×10^{-3}	0.02	0.08
sex	3.0×10^{-3}	0.01	6.0×10^{-3}	2.0×10^{-3}	9.0×10^{-3}
familial MBCNs	0.02	0.04	3.0×10^{-3}	0.18	0.07
familial cases other than MBCNs	5.2×10^{-4}	3.0×10^{-3}	5.7×10^{-4}	0.01	0.04
non-familial MBCNs	2.4×10^{-5}	1.5×10^{-5}	1.3×10^{-4}	4.8×10^{-5}	1.5×10^{-4}
non-familial cases other than MBCNs	2.0×10^{-3}	4.2×10^{-4}	3.0×10^{-3}	2.0×10^{-3}	3.2×10^{-4}
% trait variance accounted for by model	10.87%	10.58%	10.57%	10.45%	10.05%

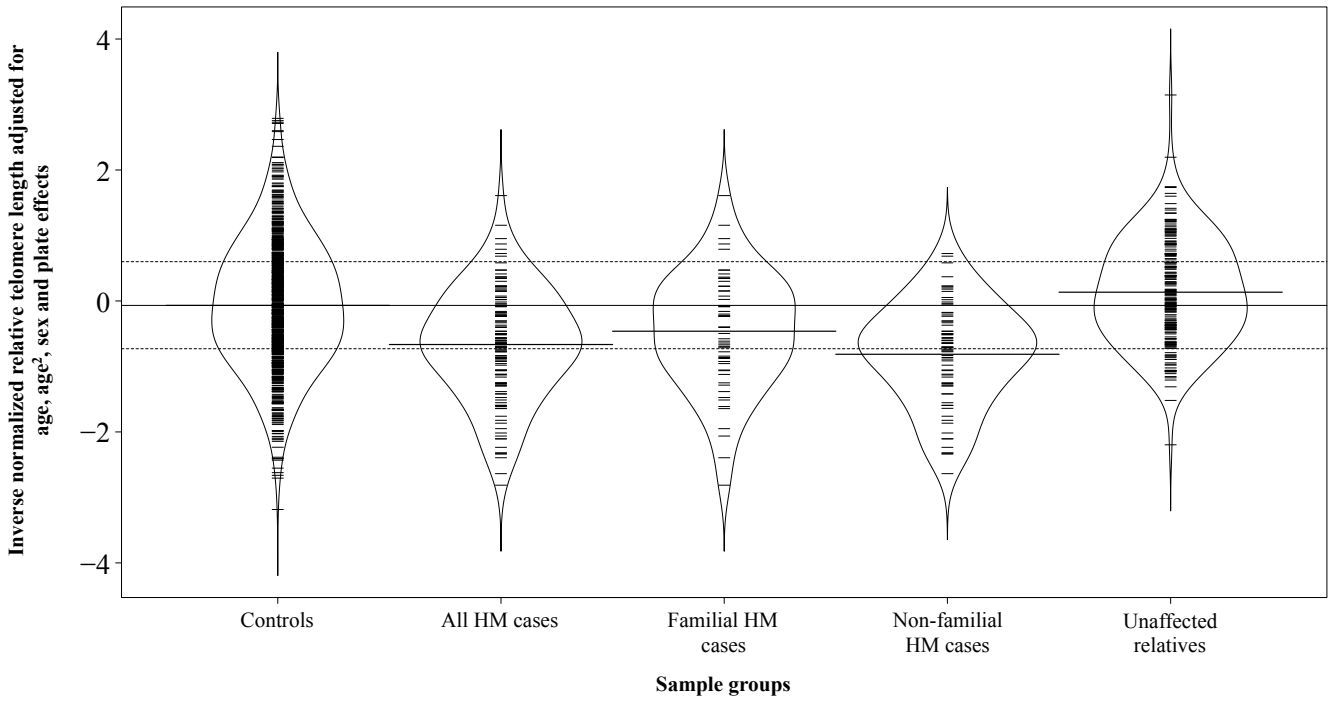


Figure 5.1 Bean plot quartile analysis of adjusted inverse normalised relative telomere lengths.
The adjusted inverse normalised relative telomere length for each group is displayed as a bean plot with individual sample measurements as lines within the bean and the overall distribution of all samples in each group shown. Horizontal bars for each bean indicate the mean of each group. The solid line and dashed lines show the mean and interquartile range of the control group.

5.5.4 Genetic variants in telomere biology genes in individuals in the lowest quartile of telomere length

Of the 31 individuals genome or exome sequenced as per Chapter 3, 21 also have telomere length measurements. Within these 21 individuals, five HM cases (LK0124-179, LK0124-202, LK0153-004, LK2042-006, LK2042-257) and one unaffected first-degree relative of a case (LK2042-018), all genome sequenced, had telomere lengths in the lowest quartile.

Overall, 145 variants were identified in 99 telomere biology genes. Table 5.5 shows the number of variants in telomere biology genes with CADD phred-like scaled C scores ≥ 10 per individual, for whom telomere length was measured. A comparison of average variant counts between individuals in the lowest quartile of telomere length and individuals not in the lowest quartile did not reveal a statistical difference either with all individuals (P=0.9), or just restricted to genome sequenced individuals (P=0.2). Restricting the variants to only telomere biology genes as specified in Mirabello *et al*[318] did not alter this result.

Table 5.5 Variants in telomere biology genes from TFHMS NGS samples, with measured telomere length, with CADD phred-like scaled C scores ≥ 10.

Individual	NGS	Lowest Quartile of Telomere length	Number variants with CADD phred-like scaled C scores ≥ 10
LK0051-001	WGS		35
LK0051-007	WES		30
LK0051-128	WGS		36
LK0051-159	WGS		41
LK0124-117	WGS		34
LK0124-179	WGS	Yes	36
LK0124-202	WGS	Yes	36
LK0139-001	WGS		39
LK0139-004	WES		29
LK0139-005	WGS		37
LK0153-003	WGS		48
LK0153-004	WGS	Yes	41
LK2042-003	WGS		40
LK2042-006	WGS	Yes	36
LK2042-018	WGS	Yes	41
LK2042-231	WGS		41
LK2042-232	WES		31
LK2042-257	WGS	Yes	32
LK2042-258	WES		29
LK2042-281	WGS		40
LK2042-290	WGS		49

5.6 Discussion

These analyses determined that mean relative telomere length is highly heritable within the TFHMS families, supporting previously reported heritability estimates in non-disease families[301-303]. The finding that mean relative telomere length was shorter in both familial and non-familial HM cases indicates that telomere length is likely to be important in the genetic etiology of HMs. A previous study of mean relative telomere length in familial myelodysplastic syndrome MDS-AML has shown that affected individuals from four small families had shorter telomeres concurrent with mutations in the telomerase gene *TERT* and its RNA component *TERC*[147]. Of the five cases across the four families reported to have shorter telomeres two had aplastic anaemia, two had MDS and one had MDS-AML. This study extends the findings of Kirwan and colleagues[147] to that of large families with multiple HM subtypes, finding new evidence of the involvement of telomere length in both familial and non-familial HMs.

Age, sex and age^2 as covariates explained a proportion of the variation in mean relative telomere length in this study. This is in keeping with telomere length declining with age and males having shorter telomeres than females[297]. A significant age^2 indicates a non-linear component in the age-related telomere length decline, a finding in line with a recent report showing a differential rate of decrease of telomere length over different age ranges[319]. The population controls measured do have a higher percentage of males, which could be suggested to be driving the association with sex, however SOLAR was used to correct mean relative telomere length for sex effects.

An important caveat with this retrospective study is that the finding of shorter mean relative telomere lengths in HM cases could also be related to disease susceptibility, treatment or the disease process. This study does not have the necessary clinical information to appropriately analyse these factors. Currently the literature surrounding the role of chemotherapeutic agents in telomere shortening remains controversial and inconclusive. In a study of telomere length in breast cancer patients before and after chemotherapy, Schröder *et al.* report individual changes in patient

telomere length, some shortening, some lengthening, but overall no statistically significant changes in response to chemotherapy[320]. Similarly in 2010 Mirabello *et al.* did not find a difference in telomere length measured in peripheral blood samples between breast cancer patients who had undergone chemotherapy and those that had not[321]. In their 1998 study of children with acute leukaemia, Engelhardt *et al.*[322] reported telomere length shortening after repeated rounds of chemotherapy but in their 2003 follow up study[323] their findings suggest that there is little effect of chemotherapy on telomere length in children with acute leukaemia and a heterogeneous effect on telomere length in children with solid tumours, similar to the findings of Schröder *et al.* in 2001[320]. In their study of APML patients, Ghaffari and colleagues showed that after chemotherapy, when patients had achieved remission, there was no difference in telomere length between patients and controls[293]. This is consistent with a study of CLL patients who were followed over time and showed that chemotherapy did not have an effect on telomere length[296]. A study of *TP53* mutation carriers in Li-Fraumeni syndrome families also found that cancer therapy is unlikely to affect telomere length[324]. Recently a small study of patients with HMs suggested that chemotherapy induces telomere length shortening[325]. However, since this study did not report pre-chemotherapy telomere length, an alternative interpretation is that telomere length shortening is a disease related feature and not chemotherapy induced[325].

It could be concluded from these reports, and others, that chemotherapy has no consistent influence on telomere length in blood cells particularly when examining multiple chemotherapeutic treatment regimens. As this study features a variety of HM subtypes in several generations exposed to a range of therapeutic regimens, analysis stratified on treatment type is not possible due to low sample numbers in each category.

A second consideration is that shorter telomeres in HM cases could be the result of malignant cell DNA within the genomic DNA sample. It is recognised that circulating malignant cells can be present for many years in chronic HM subtypes such as CLL. Based on the clinical diagnoses of HM cases in this study two additional sub-analyses of the primary data were conducted. In the first analysis all CLL cases (N=24) were removed on the basis that DNA obtained from blood of cases with this subtype of HM

was likely to contain DNA from diseased cells (Table 5.4). In the second analysis all cases for which blood samples were obtained for DNA within 2 years of diagnosis were removed (N=24; Table 5.4). Repeating the variance components modelling in these two analyses maintained the key significant associations with HM disease, suggesting that circulating disease is not contributing to the telomere length associations identified. In an additional sub-analysis all HM cases, controls and unaffected relatives (N=126) aged 80 years and above were excluded on the basis that the population HM risk increases with age. This did not change the principle findings of shorter telomeres in familial and non-familial HM cases, nor did a final combined sub-analysis excluding individuals from all 3 sub-analyses. All cases, controls and unaffected relatives were included in the primary analysis models reported in Table 5.4.

With NGS data available for 21 individuals with telomere length measurements an analysis was conducted of variants located in telomere biology genes and predicted by CADD[244] to be within the top 10% most deleterious variants. Six of the 21 individuals with NGS data available were located in the lowest quartile of mean relative telomere length (Figure 5.1), however these individuals did not show an enrichment of deleterious variants in telomere biology genes in comparison to the remaining individuals not in the lowest quartile of mean relative telomere length. This analysis was limited by the small number of individuals for whom both NGS variants and telomere length measurements were available for. Association testing using a larger number of NGS individuals comparing variants in telomere biology genes between those in the lowest and highest quartiles of telomere length, particularly comparing HM cases to population controls, is warranted given the association identified between HMs and shorter telomere length. Additionally, GWASs have identified 31 loci significantly associated with variation in telomere length[326-336], both in recognised telomere biology genes such as *TERT* and *TERC*, and genes not previously implicated in telomere length. Therefore while this analysis focussed on rare variants with MAF $\leq$ 1%, common genetic variants may also play an important role in telomere length variation and influence disease in this cohort. Further there has been interest recently in the inheritance of *TERT* promoter mutations in the predisposition to familial and sporadic melanoma[337] and other cancers[338], however in this study no such promoter

mutations were identified in the analyses. The underlying genetic factors contributing to the association between HMs and shorter telomere length remain to be identified.

5.7 Conclusion

These analyses show for the first time that mean relative telomere length is heritable in large HM families with multiple generations affected by multiple subtypes of HMs, indicating a strong genetic effect driving trait variation. Both familial and non-familial HM cases from the same population have shorter mean relative telomere length. Taken together, the results from this retrospective study provide new evidence that telomere length is an important genetic factor in a wide range of HM subtypes and in individuals with and without a family history of disease. These findings contribute further support to the use of telomere length as a prognostic risk factor for HMs. Indeed HMs may fall broadly into the telomere syndromes proposed by Armanios and Blackburn[339,340]. Interestingly, evidence for anticipation has been previously reported in Tasmanian families in this TFHM study[97] and recent studies of familial breast cancer[157], Lynch syndrome[341], dyskeratosis congenita[154] and familial chronic myeloproliferative disorders[342] have suggested that changes to telomere length may underlie the observation of anticipation. Whilst the role of telomere length in anticipation in familial HMs remains to be elucidated, the finding that telomere biology is likely to be involved provides clues as to the underlying genetic mechanism of disease.

Chapter 6 - Conclusions

6.1 A strong but largely unknown genetic component to HMs

For haematological malignancies both population-based and family-based studies indicate a strong genetic component contributing to disease risk, with relatives of cases repeatedly shown to have an increased risk of developing disease. Further, families have been identified with clusters of HMs occurring indicating disease inheritance. The identification of a familial risk for a disease indicates that there is a genetic component underlying the disease. Identification of what specifically this genetic component is, and accordingly the cause of the familial risk, is needed to increase the understanding of the factors that lead to the development of HMs. Identification of these factors informs clinical targets for predicting an individuals risk of disease, and will ideally lead to the development of new biological markers for more efficient disease diagnosis as well as new therapeutic targets for better prognosis and patient outcomes.

Previous work toward identifying the genetic susceptibilities underlying HMs has focussed on population-based case-control studies of specific HM subtypes using genome-wide scans of common variation to identify localisation signals that are associated with increases in disease risk, as discussed in section 1.5.3. Such studies have identified and replicated a large number of common genetic variants with small increases in disease risk, which together explain only a minor portion of the inherited susceptibility for disease. While such studies have been moderately successful and contribute to the understanding of HMs, the clinical translation of these associations has yet to be realised.

The use of multigenerational families with clusters of HM cases to identify factors shared between related cases was successful in identifying the early genes contributing to disease in high risk families, as described in sections 1.5.2, 1.5.4, 1.5.5 and 1.6.2. This family-based approach has arguably shown more success than population-based approaches with several genes identified, that when mutated are a

primary causal factor in the development of disease. One classical example is the contribution of germline mutations in *RUNX1* to the development of familial AML, discussed specifically in section 1.5.2.1. While germline mutation carriers require additional somatic mutations for malignancy development, the inherited mutations in *RUNX1* are the cancer-initiating event. *RUNX1* has been shown to be dysfunctional in several cancers and there is much ongoing research into the potential of *RUNX1* directed therapies[343]. Further, the relatively recent advent of NGS and the associated bioinformatics techniques has enabled the examination of the genetics of familial HM cases in unprecedented detail and with a much sharper focus. Amongst several others, mutations in genes including *TP53*[187] and *PAX5*[189] as contributing to familial ALL, as described in section 1.6.2. While specific therapies targeted at these genes for use in mutation carriers are developing, an immediate benefit to their identification is the ability to screen these genes for mutations in other ALL patients. This is clinically significant as it can inform treatment strategies such as selection of relatives without the mutation as potential bone marrow donors for haematopoietic stem cell transplant, as has occurred in an ALL family carrying *TP53* mutations[187].

6.2 Studying HM families to identify new genetic susceptibilities

Studying HM families has been a successful approach to identifying key genetic susceptibilities underlying HMs. While the contribution of several genes to HM predisposition have been described there are families for which there is a clear genetic predisposition to HMs, based on the pattern of disease occurrence, where the underlying genetic cause has not been identified. Such families present an opportunity for identifying unknown predisposing genetic factors contributing to HM development. This is due to related affected individuals having a shared genetic background contributing to disease predisposition, which can be targeted and used to reveal new disease related factors. Identification of new disease related factors from family-based studies stands to generate new focuses for targeted therapies as a result of the increase in biological understanding of disease development. This in turn may provide new targets for drug design directed against novel cellular pathways for which the contributions to the development of HMs were previously unclear.

Accordingly, as outlined in section 1.8,we employed these newly available tools to *identify new genetic factors contributing to the familial predisposition of HMs*. To achieve this aim a rare HM family resource, the Tasmanian Familial Haematological Malignancies Study was used, as described in Chapter 2. This resource consists of large multigenerational families from a genetically homogenous Tasmanian population with clusters of multiple HM subtypes occurring in each family. Where previous studies of familial HMs have typically focussed on families with single HM subtypes for disease gene identification, the TFHMS families instead show inheritance of multiple subtypes of HMs. This is an important feature of this resource as it allows identification and exploration of the contribution of genetic factors to HM predisposition in families with mixed disease phenotypes. The concept of a shared genetic predisposition to multiple HM subtypes is intuitive when haematopoietic stem cell differentiation is considered as a continuum, whereby a mutation may predispose to multiple malignancies across the spectrum of haematopoiesis, as discussed in section 1.2.

6.3 Application of next generation sequencing to the TFHMS to identify variants predisposing to familial HMs

To identify new genetic factors contributing to familial HM predisposition a NGS approach using both genome and exome sequencing was used. As described in Chapter 2, five families from the TFHMS were selected for genome and exome sequencing, with the primary aim of identifying novel inherited variants that are likely to contribute to HM disease predisposition. Families LK0051, LK0139 and LK0153 had dense aggregations of HMs occurring in first and second-degree relatives. Family LK0124, comprising three sequenced cases, had evidence of clinically interesting HM subtypes with rare CNS presentations, with the cases selected for sequencing including a family member with a CNS lymphoma. Family LK2042 was the largest family studied with eight HM cases selected for sequencing. As for family LK0124 these cases were spread across multiple branches of the family and were selected to address the question of whether distantly related cases, separated by eight meioses or more, have a shared genetic predisposition to disease. In total 31 individuals were sequenced, including 18 HM cases and 13 unaffected relatives.

6.3.1 A tiered prioritisation analysis strategy for variant identification

To analyse and identify rare disease related variants from the NGS data a tiered prioritisation analysis strategy was developed in Chapter 3, as specifically described in section 3.3.7. This strategy made use of family specific sharing analyses to identify variants shared amongst related cases but not unaffected, non-obligate carrier relatives. This approach was used because identifying variants shared by related cases increases the evidence that the variant is disease causing, and reduces the spectrum of variants to interrogate. Importantly the tiered prioritisation strategy used both variant-based and gene-based evidence to identify new candidate variants contributing to disease. It was reasoned that variants that had a bioinformatics prediction of having an effect on gene function, together with evidence of involvement of the gene in phenotypes related to HMs and the known genetic basis of HMs were more likely to be related to HM predisposition and therefore were a higher priority for analysis.

Together the outcome of this strategy was the identification of 26 novel candidate HM susceptibility variants in 26 genes across the five families, as described in section 3.4.4. The identification of multiple variants likely to be related to disease in each family differs from previous studies of familial HMs as described in section 1.5 and 1.6.2. Previous studies have focussed on specific genetic linkage regions or have been biased studies of specific candidate genes[148,191]. Both types of study design have potentially missed variants contributing to disease that were not in their focus regions or candidate genes. Here the unbiased analysis used instead identified multiple variants in each study family, which is in line with the concept of multiple genetic factors contributing to complex diseases such as HMs.

6.3.2 Variants in *TDP2* and *TNFSF9* implicated in HM predisposition

The two variants from family LK0051 with the most compelling biological evidence of a likely involvement in the development of HMs were validated by Sanger sequencing and prioritised to establish further evidence of their contribution to HM predisposition. These variants were rs200729372 in *TDP2* and rs61750000 in *TNFSF9*. The aim here was to identify whether the identified variants were enriched in the TFHMS resource in comparison to population controls using high-throughput genotyping. As described in section 4.4.3 these analyses found that the *TDP2* variant

was significantly enriched in the TFHMS resource, in comparison to the Exome Variant Server control data (P=0.0028), and the *TNFSF9* variant as significantly enriched in comparison to both Tasmanian population controls (P=0.03), and EVS control data (P=0.01). The identification of recurrence of these variants in other familial and non-familial HM cases is important evidence supporting their contribution to disease. This is in line with other familial HM studies that have used variant recurrence as a measure of the contribution of the variant to disease risk[148,189] with recurring variants in multiple HM families, as is seen here, being compelling support that the variants are involved in HM susceptibility.

Additionally, other familial HM studies have used whole gene sequencing in sets of familial and non-familial HM cases to identify additional variants in the implicated gene contributing to disease[148]. In this dissertation, targeted NGS of the *TNFSF9* gene in 96 additional HM cases did not identify further variants related to disease as described in section 4.4.4, with the rs61750000 variant remaining the most biologically significant variant identified. However screening *TNFSF9* within a larger collection of HM samples, by accessing a tissue bank such as the Australasian Leukaemia & Lymphoma Group[344], specifically HM cases with similar subtypes to the variant carriers in this study, may identify other variants in this gene involved in disease susceptibility.

Within the LK0051 family the *TNFSF9* variant occurred in the WGS uncle and nephew pair, who were affected by a T-cell lymphoma, and a B-cell Burkitt lymphoma respectively, as well as their connecting unaffected female relative. The occurrence of disease variants in unaffected family members has been found in other studies[147,148,189] and is expected as the underlying premise of identifying these variants is that they contribute to disease predisposition with additional genetic changes required for malignancy to develop. In contrast to previous studies that have focussed on identifying predisposition variants in single HM subtypes, *a novel finding from this study was the identification and validation of variants shared in multiple HM subtypes in the same family*. This is supportive of a common genetic predisposition to multiple subtypes and is in line with the continuum of haematopoietic differentiation discussed in section 1.2 whereby inherited mutations predispose to multiple different types of HMs.

A further eleven variants in three families were validated by Sanger sequencing with each warranting TFHMS resource and control population based screening to identify whether these variants are enriched in HM families and potentially contributing to disease susceptibility.

6.4 Future directions to confirm a biological role for *TNFSF9* in HM susceptibility

6.4.1 Previous research supports a role for *TNFSF9* in the biology of HMs

Previous research strongly supports a role for *TNFSF9* in the biology of HM development in both B-cell and T-cell malignancies. When this gene is deleted in mice, 60% of them develop B-cell lymphomas within a year of birth[281]. It is recurrently deleted in human HMs in cases of DLBCL and BL[182,280] and has established biological roles in both B-cell, and T-cell, proliferation and function[273,278,279]. Examining *TNFSF9* through the cBioPortal for Cancer Genomics (January 20, 2015 release), an open-access resource for exploring cancer genomics datasets[345,346] reveals no somatic cancer genome mutations or copy number alterations to *TNFSF9* in HM datasets, however a recurring deletion mutation L41del is present in 19 of 90 pancreatic cancer cases indicating that *TNFSF9* may also play a role in subset of pancreatic cancers. A recent study of 166 B-cell lymphomas with bone marrow involvement has shown expression of *TNFSF9* in greater than 50% of samples studied and proposes that *TNFSF9* expression be used as a factor in HM disease staging as well as minimal residual disease assessment[347]. Further, *TNFSF9* in this study was also found to differentiate bone marrow involvement of B-cell lymphomas from normal and reactive lymphoid infiltrates[347]. Recent work has also shown that in addition to its roles in lymphoid cell biology, *TNFSF9* also has a role in myeloid cell biology. While Zhao and colleagues found that *TNFSF9* expression is absent in myeloid cells as well as AML[347], Cheng and colleagues found that activation of *TNFSF9* through application of recombinant TNFRSF9, the receptor for TNFSF9's cell-to-cell interactions with T-cells, induced differentiation of malignant cells in AML patients[348]. Together these studies provide further evidence supporting a

role for disruption in *TNFSF9* through inherited variants contributing to HM development.

6.4.2 Determining the function of the *TNFSF9* rs61750000 variant

As the identified variant is a coding, non-regulatory change, the analysis of gene expression does not immediately present itself as an ideal strategy to show an effect on gene function. However, potentially, variant carriers may have an increase in *TNFSF9* expression in an attempt to compensate for a deficiency in gene function resulting from the variant. Alternatively, if the variant results in altered *TNFSF9* activity there may be corresponding changes in downstream gene expression that could be detectable in variant carriers. It is possible that both loss of function, as observed by somatic *TNFSF9* deletions in HMs, and gain or increase of function through mutation or overexpression can paradoxically both contribute to malignancy development. Evidence suggests that it is disruption of *TNFSF9* function that contributes to HMs with the direction of the effect of the disruption (gain or loss) potentially modulating the size of the contribution. Within the TFHMS study, as RNA is available for both the HM affected mutation carriers and several TFHMS cases without this variant, a logical starting point would be to examine *TNFSF9* gene expression.

With compelling support for a role for the *TNFSF9* gene in HM predisposition, the challenge is then to determine whether the specific variant, rs6175000, causing the amino acid change G139A, is deleterious to protein function. Contributing to its identification in Chapter 3 and prioritisation for further experimentation in Chapter 4 of this study, at the *in silico* level bioinformatics tools, including CADD[244], PolyPhen-2[349] and SIFT[350], consistently predict that this variant has a functional effect on the TNFSF9 protein. However specific experiments are required that demonstrate a direct effect of this variant on the biological role of *TNFSF9*.

For future work to demonstrate this effect, cell culture models of B-cells and T-cells present as an attractive option for addressing whether the *TNFSF9* G193A variant is biologically important. Cell lines could be altered to overexpress the TNFSF9 G139A variant, and the native TNFSF9 protein, and the effect on cell proliferation and

survival measured comparatively. Alternatively, induced pluripotent stem cell models could be developed from living variant carriers, allowing targeted differentiation and direct characterisation of the biology of a model of haematopoietic cells of variant carriers and exploration of the effect of correcting the variant using genome editing technologies such as the CRISPR/Cas9 system, as has been applied in a recent study of β-thalassemia [351]. The results from these experiments would show whether the variant has a functional effect on T-cell and B-cell biology. Evidence of a biological effect of the variant should be followed by protein centred experiments to determine the mechanism by which the G193A variant exerts an effect.

Potentially, protein-centred future work could use protein modelling and protein-protein interaction experiments to determine the effect of the G193A variant on the ability of the TNFSF9 trimer to form, or the ability of the TNFSF9 trimer to bind with its TNFRSF9 receptor. Together experiments targeted at determining the biological role of the G193A variant in *TNFSF9* in contributing to HM predisposition, together with experiments targeted at the mechanism of how this variant affects gene function are needed to provide the necessary evidence for disease causality.

6.5 Telomere length is a potential risk factor for haematological malignancies

Concurrent with the identification of shared variants likely to contribute to HM predisposition another major finding of this dissertation was an association between telomere length and HMs. This work was published in Blackburn et al.[1]. Telomere biology has a well-characterised role in cancer development. Disruption of key telomere biology genes has been shown to lead to a spectrum of syndromes, of which haematological malignancies are a feature, such as dyskeratosis congentia and aplastic anaemia[340]. Further, mutations in key telomere biology genes have been identified in families with MDS/AML[147]. Therefore, the aim of this project was to measure telomere length, a quantitative trait relevant to cancer development, across the TFHMS resource to identify new evidence implicating telomere length as a genetic factor in HMs. The findings from these analyses showed that telomere length, as a trait is highly heritable (62.5%, $P=4.7\times10^{-5}$), and shorter in HM cases ($P=2.9\times10^{-6}$), with both familial and non-familial cases having shorter telomeres ($P=2.2\times10^{-4}$ and

P=2.2×10^{-5} respectively). Several analyses were conducted to determine whether the observed association with HM case status and shorter telomere length was explained by particular sub-divisions of the cases, based on familial and non-familial cases or major HM types. *Regardless of how HM cases were analysed results consistently showed shorter telomeres in HM cases, in line with the hypothesis that telomere length is a risk factor for HMs.* However, as a retrospective study, accounting for factors such as disease duration, chemotherapy and radiotherapy, was not possible in these analyses. These findings do highlight that telomere length is a potential risk factor in HMs, suggesting that larger prospective studies of populations measuring telomere length and HM incidence over time are warranted. Further, a wide variation of telomere length was observed in the population controls used in this study. This together with the observed association with HMs, suggests that telomere length as a clinical measurement may be of use in conjunction with other disease measurements or genetic markers as a method of stratifying existing HM cases and predicting treatment outcomes, rather than a population level screen to determine HM risk.

6.6 Recommendations for future familial HM studies

The advances to genetics research provided by NGS have renewed interest in the potential of family-based resources, such as the TFHMS, for the identification of the genetic variants underlying inherited disease and the work presented here in this dissertation can be used to suggest a number of recommendations to help guide genetic studies utilising a similar family-based approach.

6.6.1 Selection of family members for sequencing

The strength of familial resources is tied to their family pedigrees and sample availability. Identification and recruitment of families with clusters of HM cases separated by no more than five meioses (second cousins), or more distantly related cases with a clear genetic transmission of the disease, is essential. Recruitment and sequencing of unaffected relatives is also informative as variants identified in these individuals can be used as within and between family controls, as the HM unaffected branch of the LK0153 family was used in this study in Chapter 3. When selecting unaffected family members for sequencing, ideally they should undergo standard

clinical evaluation to determine whether they are disease-free or potentially have a pre-malignant condition. Such comprehensive clinical evaluation is not always possible or practical, therefore selection of older unaffected relatives, particularly when the HM cases in that family have an earlier than average onset of disease, is a valuable alternative. Sequencing of unaffected relatives as within family controls must be done with due consideration to disease penetrance. In HMs, due to the multifactorial nature of disease, unaffected relatives may be disease-free but carry some or all of the disease predisposing variants present in the related HM cases. In this study, when unaffected siblings were sequenced, such as in family LK0139, variants within the unaffected sibling were removed from the HM cases in the heuristic filtering analysis described in section 3.3.7. This is one valid strategy; an alternative would be to not filter these variants, at the cost of keeping variants not connected to disease and consequently more candidate variants to prioritise. Presence or absence of variants in the unaffected relative could then be informative in the final stage of variant prioritisation after other possibilities have been filtered.

6.6.2 Focus on families with closely related HM cases

Overall, in this study, sequencing of closely related clusters of HM cases had the most success in identifying shared candidate variants that validated when analysed using Sanger sequencing as described in section 4.4.1. While the variants identified in family LK2042, the largest family with multiple distantly related HM cases sequenced, were not selected for validation by Sanger sequencing, those identified as shared between cases in the LK0124 family presented as a comparable family structure for evaluation of sequencing distantly related cases. Indeed the results of the sharing based analyses of family LK0124, which identified three candidate variants, did not validate when these variants were Sanger sequenced. Two of these variants that failed validation were also identified as false positives in other sequenced families.

These findings from the analyses in Chapter 3 and 4 highlight two points that may be useful in future analyses of similar families. Firstly, rare variants that are identified across multiple families may be false positives generated as an artefact of the sequencing, alignment or variant calling processes. This is useful to know as such

variants can be scrutinised closer during prioritisation. The second point is that studies such as this should focus on sequencing closely related HM cases, within five meioses of each other, as rare variants shared by distant relatives that are separated by eight meioses or greater, are unlikely to be true variants when using the heuristic filtering approach employed in this study, as was shown for family LK0124 in section 4.4.1.

To that end a future direction for NGS in the TFHMS, would be to focus on two recently identified families in which newly diagnosed HM cases were identified at the end of this dissertation. One is a family of five siblings with three affected by B-cell HMs and two unaffected siblings, all of similar ages. The other is an affected sibling pair with an affected first cousin, all with B-cell HMs. Until recently each of these families contained only two HM cases, however continued genealogical monitoring of HM incidence using the Tasmanian Cancer Registry identified the third case in each family when diagnosed, highlighting the importance of a longitudinal familial resource such as the TFHMS.

Disease subtype is an additional important consideration in selection of families for NGS based studies. The focus in this field has been primarily on studies of families with single subtypes of HMs, which has successfully identified disease related mutations in genes such as *PAX5* in familial ALL[189]. A valuable and important aspect of this dissertation was that the families studied had multiple different subtypes of HMs. From Chapters 3 and 4 *a novel finding was that disease related variants were shared between relatives with different HM subtypes, supporting a shared predisposition to multiple subtypes.* Future NGS studies should therefore not be limited to single HM subtype families, in contrast to the current families studied by NGS and reported in the literature.

6.6.3 Recent expert guideline recommendations for implicating causality and identifying cancer predisposition genes

Recently, expert perspective guidelines have been published on the process of investigating causality of identified NGS variants in human disease[352], and on identifying familial cancer predisposition genes[111]. MacArthur *et al.*[352] highlight that implicating a gene and variant in a disease is a process involving the accumulation and integration of different levels of evidence that together support a role for the gene

or variant in disease. Evidence is required at both the variant and the gene level. Bioinformatics tools should predict that the variant has a deleterious effect on gene function and this prediction must be confirmed experimentally. The gene itself should be expressed in tissues relevant to the disease, and demonstrated in model systems such as transgenic mice or cell culture that when disrupted a phenotype consistent with the disease is generated.

The process for disease variant identification used in this study was underpinned by layers of biological evidence implemented through a tiered prioritisation strategy that lead to the identification of the thirteen validated sequence variants in three TFHMS families. In line with the guidelines from MacArthur *et al.*[352] at the variant-level bioinformatics tools predict that the top candidate variant from this study, G193A in *TNFSF9*, affects gene function. At the gene-level, this gene is expressed in haematopoietic tissues and a HM phenotype has been demonstrated in transgenic mice where the function of the gene has been disrupted. Additionally the gene has been shown in humans to be a HM disease marker and is intrinsically involved in the cell biology of HM development.

Nazneen Rahman, a leading researcher in the field of familial cancer, in her 2014 review[111], cautions that many reported cancer predisposition genes are incorrectly classified due to mistaken data interpretations including overestimation of causality due to a variant being rare, coding and not in a control population. Collectively, rare coding variants occur frequently[353], and studying families inherently increases the rate at which rare coding variants are identified in related cases due to multiple copies of the variants being present in the family. Rahman proposes that support for a suggested cancer predisposition gene must stem from identification of multiple families where this gene is dysfunctional, requiring large-scale international collaborations. In light of this recommendation, the *TNFSF9* variant, which was identified as present in multiple TFHMS families and enriched in the TFHMS resource in comparison to population controls, remains a striking candidate for HM predisposition and further investigation.

6.8 Conclusion

The field of NGS is still in its infancy. Each year, with falling sequencing costs, development of new data sources and new analysis methodologies, this area is dynamic and ever changing. In the context of familial HMs the application of this technology to multigenerational families with multiple subtypes of disease and clusters of HM cases stands to identify the predisposing genetic factors underlying disease. In Chapter 3 a tiered prioritisation strategy was applied to NGS data from five Tasmanian families with HMs, identifying twenty-six likely candidate variants contributing to disease predisposition. In Chapter 4 thirteen of these variants were validated in three of these families. These variants are new candidates in the field of HM susceptibility, particularly the rs61750000 variant in *TNFSF9*, warranting further analysis as to their contributions to disease. These variants may lead to identification of new pathways and drug targets for the diagnosis and management of HMs. Indeed recent research in *TNFSF9* implicating it as a biomarker in HMs highlights the potential translation of these findings in the broader HM context. In Chapter 5 new evidence supporting a role for telomere length as a genetic factor in HMs was identified and indicates that telomere biology should be examined closer in this disease as telomere length may also be a prognostic factor clinically useful for diagnosis, disease risk estimation, or disease monitoring in HMs.

Novel genetic factors were identified in each family analysed by NGS in this study. These findings show that there appears to be multiple rare variants predicted to impact on gene function that are likely to contribute to HM susceptibility. This serves to highlight the challenges that are being encountered utilising these approaches. This is consistent with similar published studies and further highlights the complexity of the genetic architecture underlying the predisposition to HMs.

References

1. Blackburn, N. B. *et al.* A retrospective examination of mean relative telomere length in the Tasmanian Familial Hematological Malignancies Study. *Oncol Rep* **33,** 25–32 (2015).
2. Australian Institute of Health and Welfare & Australasian Association of Cancer Registries 2012. Cancer in Australia: an overview, 2012. 1–215 (2012). at <https://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129542353>
3. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. (2013). at <http://globocan.iarc.fr>
4. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
5. Weinberg, R. *The Biology of Cancer, Second Edition*. (Garland Science, 2013).
6. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
7. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
8. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132,** 631–644 (2008).
9. Hatton, C. S. R., Hughes-Jones, N. C., Hay, D. & Keeling, D. *Lecture Notes: Haematology*. (John Wiley & Sons, 2013).
10. Ceredig, R., Rolink, A. G. & Brown, G. Models of haematopoiesis: seeing the wood for the trees. *Nat. Rev. Immunol.* **9,** 293–300 (2009).
11. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117,** 663–676 (2004).
12. Laiosa, C. V., Stadtfeld, M., Xie, H., de Andres-Aguayo, L. & Graf, T. Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBP alpha and PU.1 transcription factors. *Immunity* **25,** 731–744 (2006).
13. Iwasaki, H. & Akashi, K. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26,** 726–740 (2007).
14. Sudo, K., Ema, H., Morita, Y. & Nakauchi, H. Age-associated characteristics of murine hematopoietic stem cells. *J. Exp. Med.* **192,** 1273–1280 (2000).
15. Rossi, D. J. *et al.* Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc Natl Acad Sci USA* **102,** 9194–9199 (2005).
16. Rossi, D. J., Bryder, D., Seita, J. & Nussenzweig, A. Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature* (2007).
17. Nijnik, A., Woodbine, L., Marchetti, C., Dawson, S. & Lambe, T. DNA repair is limiting for haematopoietic stem cells during ageing. *Nature* (2007).
18. Okuda, T., Nishimura, M., Nakao, M. & Fujita, Y. RUNX1/AML1: a central player in hematopoiesis. *Int. J. Hematol.* **74,** 252–257 (2001).
19. Lutterbach, B. & Hiebert, S. W. Role of the transcription factor AML-1 in

acute leukemia and hematopoietic differentiation. *Gene* **245,** 223–235 (2000).
20. Harada, H. *et al.* High incidence of somatic mutations in the AML1/RUNX1 gene in myelodysplastic syndrome and low blast percentage myeloid leukemia with myelodysplasia. *Blood* **103,** 2316–2324 (2004).
21. Ellinghaus, E. *et al.* Identification of germline susceptibility loci in ETV6-RUNX1-rearranged childhood acute lymphoblastic leukemia. *Leukemia* **26,** 902–909 (2012).
22. Greene, M. E. *et al.* Mutations in GATA1 in both transient myeloproliferative disorder and acute megakaryoblastic leukemia of Down syndrome. *Blood Cells Mol Dis* **31,** 351–356 (2003).
23. Khan, I., Malinge, S. & Crispino, J. Myeloid leukemia in down syndrome. *Crit Rev Oncog* **16,** 25–36 (2011).
24. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. *IARC: Lyon* (2008).
25. *Australian Institute of Health and Welfare*. (Australian Government). at <http://www.aihw.gov.au/cancer-data/>
26. Bray, F., Ren, J.-S., Masuyer, E. & Ferlay, J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer* **132,** 1133–1145 (2013).
27. Hawkins, M. M. Long-term survivors of childhood cancers: what knowledge have we gained? *Nat Clin Pract Oncol* **1,** 26–31 (2004).
28. Maule, M. *et al.* Risk of second malignant neoplasms after childhood leukemia and lymphoma: an international study. *J Natl Cancer Inst* **99,** 790–800 (2007).
29. Hallek, M. *et al.* Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet* **376,** 1164–1174 (2010).
30. Hallek, M. Chronic lymphocytic leukemia: 2013 update on diagnosis, risk stratification and treatment. *Am. J. Hematol.* **88,** 803–816 (2013).
31. Yang, Y., Li, T. & Nielsen, M. E. Aging and cancer mortality: dynamics of change and sex differences. *Exp. Gerontol.* **47,** 695–705 (2012).
32. DeGregori, J. Challenging the axiom: does the occurrence of oncogenic mutations truly limit cancer development with age? *Oncogene* **32,** 1869–1875 (2013).
33. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J Clin* **61,** 69–90 (2011).
34. Cook, M. B., McGlynn, K. A., Devesa, S. S., Freedman, N. D. & Anderson, W. F. Sex disparities in cancer mortality and survival. *Cancer Epidemiol Biomarkers Prev* **20,** 1629–1637 (2011).
35. Najari, B. B. *et al.* Sex disparities in cancer mortality: the risks of being a man in the United States. *J. Urol.* **189,** 1470–1474 (2013).
36. Cartwright, R. A., Gurney, K. A. & Moorman, A. V. Sex ratios and the risks of haematological malignancies. *Br J Haematol* **118,** 1071–1077 (2002).
37. Dorak, M. T. & Karpuzoglu, E. Gender differences in cancer susceptibility: an inadequately addressed issue. *Front Genet* **3,** 268 (2012).
38. Edgren, G., Liang, L., Adami, H. O. & Chang, E. T. Enigmatic sex disparities in cancer incidence. *European journal of ...* (2012).
39. World Health Organization International Agency for Research on Cancer. *World Cancer Report 2008*. 1–260 (International Agency for Research on Cancer, 2008).
40. Rosenberg, P. S., Wilson, K. L. & Anderson, W. F. Are incidence rates of

adult leukemia in the United States significantly associated with birth cohort? *Cancer Epidemiol Biomarkers Prev* **21,** 2159–2166 (2012).
41. Vineis, P. *et al.* Tobacco and cancer: recent epidemiological evidence. *J Natl Cancer Inst* **96,** 99–106 (2004).
42. Infante, P. F. Benzene exposure and multiple myeloma: a detailed meta-analysis of benzene cohort studies. *Ann N Y Acad Sci* **1076,** 90–109 (2006).
43. Costantini, A. S. *et al.* Risk of leukemia and multiple myeloma associated with exposure to benzene and other organic solvents: evidence from the Italian Multicenter Case-control study. *Am. J. Ind. Med.* **51,** 803–811 (2008).
44. Cogliano, V. J., Baan, R., Straif, K.IARC Monographs programme staff. Updating IARC's carcinogenicity assessment of benzene. *Am. J. Ind. Med.* **54,** 165–167 (2011).
45. Prise, K. M. & Saran, A. Concise Review: Stem Cell Effects in Radiation Risk. *Stem Cells* **29,** 1315–1321 (2011).
46. Nakamura, N. A hypothesis: radiation-related leukemia is mainly attributable to the small number of people who carry pre-existing clonally expanded preleukemic cells. *Radiat. Res.* **163,** 258–265 (2005).
47. Bithell, J. F. & Stewart, A. M. Pre-natal irradiation and childhood malignancy: a review of British data from the Oxford Survey. *Br J Cancer* **31,** 271–287 (1975).
48. Doll, R. & Wakeford, R. Risk of childhood cancer from fetal irradiation. *Br J Radiol* **70,** 130–139 (1997).
49. Richardson, D. *et al.* Ionizing radiation and leukemia mortality among Japanese Atomic Bomb Survivors, 1950-2000. *Radiat. Res.* **172,** 368–382 (2009).
50. Hsu, W.-L. *et al.* The incidence of leukemia, lymphoma and multiple myeloma among atomic bomb survivors: 1950-2001. *Radiat. Res.* **179,** 361–382 (2013).
51. Iwanaga, M. *et al.* Risk of myelodysplastic syndromes in people exposed to ionizing radiation: a retrospective cohort study of Nagasaki atomic bomb survivors. *J. Clin. Oncol.* **29,** 428–434 (2011).
52. Romanenko, A. Y. *et al.* The Ukrainian-American study of leukemia and related disorders among Chornobyl cleanup workers from Ukraine: III. Radiation risks. *Radiat. Res.* **170,** 711–720 (2008).
53. Kesminiene, A. *et al.* Risk of hematological malignancies among Chernobyl liquidators. *Radiat. Res.* **170,** 721–735 (2008).
54. Muirhead, C. R. *et al.* Mortality and cancer incidence following occupational radiation exposure: third analysis of the National Registry for Radiation Workers. *Br J Cancer* **100,** 206–212 (2009).
55. Krestinina, L. *et al.* Leukemia incidence among people exposed to chronic radiation from the contaminated Techa River, 1953-2005. *Radiat Environ Biophys* **49,** 195–201 (2010).
56. Tajima, K. The 4th nation-wide study of adult T-cell leukemia/lymphoma (ATL) in Japan: estimates of risk of ATL and its geographical and clinical features. The T- and B-cell Malignancy Study Group. *Int. J. Cancer* **45,** 237–243 (1990).
57. Itoyama, T. *et al.* Cytogenetic analysis and clinical significance in adult T-cell leukemia/lymphoma: a study of 50 cases from the human T-cell leukemia virus type-1 endemic area, Nagasaki. *Blood* **97,** 3612–3620 (2001).
58. Nicot, C. Current views in HTLV-I-associated adult T-cell

leukemia/lymphoma. *Am. J. Hematol.* **78,** 232–239 (2005).
59. Levine, A. M. *et al.* Evolving characteristics of AIDS-related lymphoma. *Blood* **96,** 4084–4090 (2000).
60. Knowles, D. M. Etiology and pathogenesis of AIDS-related non-Hodgkin's lymphoma. *Hematology/oncology clinics of North America* **17,** 785–820 (2003).
61. Cheung, M. C., Pantanowitz, L. & Dezube, B. J. AIDS-related malignancies: emerging challenges in the era of highly active antiretroviral therapy. *The Oncologist* **10,** 412–426 (2005).
62. Kutok, J. L. & Wang, F. Spectrum of Epstein-Barr virus-associated diseases. *Annu Rev Pathol* **1,** 375–404 (2006).
63. Molyneux, E. M. *et al.* Burkitt's lymphoma. *Lancet* **379,** 1234–1244 (2012).
64. Pang, J., Cook, L. S., Schwartz, S. M. & Weiss, N. S. Incidence of leukemia in Asian migrants to the United States and their descendants. *Cancer Causes Control* **13,** 791–795 (2002).
65. Shirley, M. H., Sayeed, S., Barnes, I., Finlayson, A. & Ali, R. Incidence of haematological malignancies by ethnic group in England, 2001-7. *Br J Haematol* **163,** 465–477 (2013).
66. Kawamata, N. *et al.* Genetic differences between Asian and Caucasian chronic lymphocytic leukemia. *Int. J. Oncol.* **43,** 561–565 (2013).
67. Videæk, A. *Heredity in Human Leukemia and Its Relation to Cancer*. (HK Lewis & Company, 1947).
68. *Swedish Cancer Registry*. (Socialstyrelsen). at <http://www.socialstyrelsen.se/register/halsodataregister/cancerregistret/inenglish>
69. Ekbom, A. The Swedish Multi-generation Register. *Methods Mol. Biol.* **675,** 215–220 (2011).
70. Goldin, L. R., Björkholm, M., Kristinsson, S. Y., Turesson, I. & Landgren, O. Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. *Haematologica* **94,** 647–653 (2009).
71. Goldin, L. R. *et al.* Familial aggregation of Hodgkin lymphoma and related tumors. *Cancer* **100,** 1902–1908 (2004).
72. Chang, E. T. *et al.* Family history of hematopoietic malignancy and risk of lymphoma. *J Natl Cancer Inst* **97,** 1466–1474 (2005).
73. Altieri, A., Bermejo, J. L. & Hemminki, K. Familial aggregation of lymphoplasmacytic lymphoma with non-Hodgkin lymphoma and other neoplasms. *Leukemia* **19,** 2342–2343 (2005).
74. Altieri, A., Bermejo, J. L. & Hemminki, K. Familial risk for non-Hodgkin lymphoma and other lymphoproliferative malignancies by histopathologic subtype: the Swedish Family-Cancer Database. *Blood* **106,** 668–672 (2005).
75. Mensah, F. K., Willett, E. V., Ansell, P., Adamson, P. J. & Roman, E. Non-Hodgkin's lymphoma and family history of hematologic malignancy. *Am J Epidemiol* **165,** 126–133 (2007).
76. Kerber, R. A. & O'Brien, E. A cohort study of cancer risk in relation to family histories of cancer in the Utah population database. *Cancer* **103,** 1906–1915 (2005).
77. Wang, S. S. *et al.* Family history of hematopoietic malignancies and risk of non-Hodgkin lymphoma (NHL): a pooled analysis of 10 211 cases and 11 905 controls from the International Lymphoma Epidemiology Consortium

(InterLymph). *Blood* **109,** 3479–3488 (2007).
78. Kristinsson, S. Y. *et al.* Risk of lymphoproliferative disorders among first-degree relatives of lymphoplasmacytic lymphoma/Waldenstrom macroglobulinemia patients: a population-based study in Sweden. *Blood* **112,** 3052–3056 (2008).
79. Landgren, O. *et al.* Increased risks of polycythemia vera, essential thrombocythemia, and myelofibrosis among 24,577 first-degree relatives of 11,039 patients with myeloproliferative neoplasms in Sweden. *Blood* **112,** 2199–2204 (2008).
80. Goldin, L. R., Björkholm, M., Kristinsson, S. Y., Turesson, I. & Landgren, O. Highly increased familial risks for specific lymphoma subtypes. *Br J Haematol* **146,** 91–94 (2009).
81. Kristinsson, S. Y. *et al.* Patterns of hematologic malignancies and solid tumors among 37,838 first-degree relatives of 13,896 patients with multiple myeloma in Sweden. *Int. J. Cancer* **125,** 2147–2150 (2009).
82. Landgren, O. *et al.* Risk of plasma cell and lymphoproliferative disorders among 14621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden. *Blood* **114,** 791–795 (2009).
83. Goldin, L. R. *et al.* Familial aggregation of acute myeloid leukemia and myelodysplastic syndromes. *J. Clin. Oncol.* **30,** 179–183 (2012).
84. Poppe, B. *et al.* Chromosomal aberrations in Bloom syndrome patients with myeloid malignancies. *Cancer Genet Cytogenet* **128,** 39–42 (2001).
85. Masmoudi, A. *et al.* Clinical and laboratory findings in 8 patients with Bloom's syndrome. *J Dermatol Case Rep* **6,** 29–33 (2012).
86. Walne, A. J. & Dokal, I. Dyskeratosis Congenita: a historical perspective. *Mech Ageing Dev* **129,** 48–59 (2008).
87. Du, H.-Y. *et al.* Complex inheritance pattern of dyskeratosis congenita in two families with 2 different mutations in the telomerase reverse transcriptase gene. *Blood* **111,** 1128–1130 (2008).
88. Alter, B. P., Giri, N., Savage, S. A. & Rosenberg, P. S. Cancer in dyskeratosis congenita. *Blood* **113,** 6549–6557 (2009).
89. Jønsson, V. *et al.* CLL family 'Pedigree 14' revisited: 1947-2004. *Leukemia* **19,** 1025–1028 (2005).
90. Rumi, E. *et al.* Familial chronic myeloproliferative disorders: clinical phenotype and evidence of disease anticipation. *J. Clin. Oncol.* **25,** 5630–5635 (2007).
91. Jain, M., Ascensao, J. & Schechter, G. P. Familial myeloma and monoclonal gammopathy: a report of eight African American families. *Am. J. Hematol.* **84,** 34–38 (2009).
92. Tegg, E. M. *et al.* Evidence for a common genetic aetiology in high-risk families with multiple haematological malignancy subtypes. *Br J Haematol* **150,** 456–462 (2010).
93. Mack, T. M. *et al.* Concordance for Hodgkin's disease in identical twins suggesting genetic susceptibility to the young-adult form of the disease. *N Engl J Med* **332,** 413–418 (1995).
94. Belson, M., Kingsley, B. & Holmes, A. Risk factors for acute leukemia in children: a review. *Environ. Health Perspect.* **115,** 138–145 (2007).
95. Gunz, F. W., Gunz, J. P., Veale, A. M., Chapman, C. J. & Houston, I. B. Familial leukaemia: a study of 909 families. *Scand J Haematol* **15,** 117–131

(1975).
96. Deshpande, H. A., Hu, X. P., Marino, P., Jan, N. A. & Wiernik, P. H. Anticipation in familial plasma cell dyscrasias. *Br J Haematol* **103,** 696–703 (1998).
97. Tegg, E. M. *et al.* Anticipation in familial hematologic malignancies. *Blood* **117,** 1308–1310 (2011).
98. Segel, G. B. & Lichtman, M. A. Familial (inherited) leukemia, lymphoma, and myeloma: an overview. *Blood Cells Mol Dis* **32,** 246–261 (2004).
99. Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75,** 1027–1038 (1993).
100. Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75,** 1215–1225 (1993).
101. Bronner, C. E. *et al.* Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* **368,** 258–261 (1994).
102. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994).
103. Nicolaides, N. C. *et al.* Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* **371,** 75–80 (1994).
104. Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263,** 1625–1629 (1994).
105. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–792 (1995).
106. Miyaki, M. *et al.* Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat Genet* **17,** 271–272 (1997).
107. The Cancer Genome Atlas. at <http://cancergenome.nih.gov/>
108. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10,** Unit 10.11–10.11.26 (2008).
109. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39,** D945–50 (2011).
110. Wellcome Trust Sanger Institute. COSMIC: Cancer Gene census. at <http://cancer.sanger.ac.uk/cancergenome/projects/census/>
111. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505,** 302–308 (2014).
112. Di Fiore, R., D'Anneo, A., Tesoriere, G. & Vento, R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J. Cell. Physiol.* **228,** 1676–1687 (2013).
113. Wolfer, A. & Ramaswamy, S. MYC and Metastasis. *Cancer Res* **71,** 2034–2037 (2011).
114. Lerman, C. & Shields, A. E. Genetic testing for cancer susceptibility: the promise and the pitfalls. *Nat Rev Cancer* **4,** 235–241 (2004).
115. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19,** 1639–1645 (2009).
116. Fitzgerald, P. H. & Hamer, J. W. Third case of chronic lymphocytic leukaemia in a carrier of the inherited Ch1 chromosome. *Br Med J* **3,** 752–754 (1969).
117. Welborn, J. Constitutional chromosome aberrations as pathogenetic events in

hematologic malignancies. *Cancer Genet Cytogenet* **149,** 137–153 (2004).
118. Salipante, S. J. *et al.* Mutations in a gene encoding a midbody kelch protein in familial and sporadic classical Hodgkin lymphoma lead to binucleated cells. *Proc Natl Acad Sci USA* **106,** 14920–14925 (2009).
119. Saarinen, S. *et al.* Analysis of KLHDC8B in familial nodular lymphocyte predominant Hodgkin lymphoma. *Br J Haematol* **154,** 413–415 (2011).
120. Ho, C. Y. *et al.* Linkage of a familial platelet disorder with a propensity to develop myeloid malignancies to human chromosome 21q22.1-22.2. *Blood* **87,** 5218–5224 (1996).
121. Song, W. J. *et al.* Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat Genet* **23,** 166–175 (1999).
122. Raval, A. *et al.* Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell* **129,** 879–890 (2007).
123. Sellick, G. S. *et al.* A high-density SNP genome-wide linkage search of 206 families identifies susceptibility loci for chronic lymphocytic leukemia. *Blood* **110,** 3326–3333 (2007).
124. Crowther-Swanepoel, D. *et al.* Genetic variation in CXCR4 and risk of chronic lymphocytic leukemia. *Blood* **114,** 4843–4846 (2009).
125. Ichikawa, M. *et al.* AML-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. *Nat Med* **10,** 299–304 (2004).
126. Bruwier, A. & Chantrain, C. F. Hematological disorders and leukemia in children with Down syndrome. *Eur. J. Pediatr.* (2011). doi:10.1007/s00431-011-1624-1
127. Buijs, A. *et al.* A novel CBFA2 single-nucleotide mutation in familial platelet disorder with propensity to develop myeloid malignancies. *Blood* **98,** 2856–2858 (2001).
128. Kirito, K. *et al.* A novel RUNX1 mutation in familial platelet disorder with propensity to develop myeloid malignancies. *Haematologica* **93,** 155–156 (2008).
129. Look, A. T. Oncogenic transcription factors in the human acute leukemias. *Science* **278,** 1059–1064 (1997).
130. Preudhomme, C. *et al.* High incidence of biallelic point mutations in the Runt domain of the AML1/PEBP2 alpha B gene in Mo acute myeloid leukemia and in myeloid malignancies with acquired trisomy 21. *Blood* **96,** 2862–2869 (2000).
131. Langabeer, S. E., Gale, R. E., Rollinson, S. J., Morgan, G. J. & Linch, D. C. Mutations of the AML1 gene in acute myeloid leukemia of FAB types M0 and M7. *Genes Chromosomes Cancer* **34,** 24–32 (2002).
132. Silva, F. P. G. *et al.* Identification of RUNX1/AML1 as a classical tumor suppressor gene. *Oncogene* **22,** 538–547 (2003).
133. Preudhomme, C. *et al.* High frequency of RUNX1 biallelic alteration in acute myeloid leukemia secondary to familial platelet disorder. *Blood* **113,** 5583–5587 (2009).
134. Sellick, G. S. *et al.* A high-density SNP genomewide linkage scan for chronic lymphocytic leukemia-susceptibility loci. *Am J Hum Genet* **77,** 420–429 (2005).
135. Broderick, P., Sellick, G., Fielding, S., Catovsky, D. & Houlston, R. Lack of a relationship between the common 18q24 variant rs12953717 and risk of

chronic lymphocytic leukemia. *Leuk Lymphoma* **49,** 271–272 (2008).
136. Fuller, S. J. *et al.* Analysis of a large multi-generational family provides insight into the genetics of chronic lymphocytic leukemia. *Br J Haematol* **142,** 238–245 (2008).
137. Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* **40,** 1204–1210 (2008).
138. Speedy, H. E. *et al.* A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* **46,** 56–+ (2014).
139. Xu, H. *et al.* Novel Susceptibility Variants at 10p12.31-12.2 for Childhood Acute Lymphoblastic Leukemia in Ethnically Diverse Populations. *J Natl Cancer Inst* (2013). doi:10.1093/jnci/djt042
140. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
141. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363,** 166–176 (2010).
142. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8,** e1000294 (2010).
143. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic Associations Created by Rare Variants Do Not Explain Most GWAS Results. *PLoS Biol* **9,** e1000579 (2011).
144. Saunders, E. J. *et al.* Fine-Mapping the HOXB Region Detects Common Variants Tagging a Rare Coding Allele: Evidence for Synthetic Association in Prostate Cancer. *PLoS Genet* **10,** e1004129 (2014).
145. Berndt, S. I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat Genet* **45,** 868–876 (2013).
146. Smith, M. L., Cavenagh, J. D., Lister, T. A. & Fitzgibbon, J. Mutation of CEBPA in familial acute myeloid leukemia. *N Engl J Med* **351,** 2403–2407 (2004).
147. Kirwan, M. J. *et al.* Defining the pathogenic role of telomerase mutations in myelodysplastic syndrome and acute myeloid leukemia. *Hum Mutat* **30,** 1567–1573 (2009).
148. Hahn, C. N. *et al.* Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet* (2011). doi:10.1038/ng.913
149. Calin, G. A. *et al.* Familial cancer associated with a polymorphism in ARLTS1. *N Engl J Med* **352,** 1667–1676 (2005).
150. Calin, G. A. *et al.* A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* **353,** 1793–1801 (2005).
151. Bullrich, F. *et al.* ATM mutations in B-cell chronic lymphocytic leukemia. *Cancer Res* **59,** 24–27 (1999).
152. Scott, R. H. *et al.* Familial T-cell non-Hodgkin lymphoma caused by biallelic MSH2 mutations. *Journal of medical genetics* **44,** e83 (2007).
153. Niemeyer, C. M. *et al.* Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat Genet* **42,** 794–800 (2010).
154. Vulliamy, T. *et al.* Disease anticipation is associated with progressive

telomere shortening in families with dyskeratosis congenita due to mutations in TERC. *Nat Genet* **36,** 447–449 (2004).
155. Armanios, M. *et al.* Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc Natl Acad Sci USA* **102,** 15960–15964 (2005).
156. Tabori, U., Nanda, S., Druker, H., Lees, J. & Malkin, D. Younger age of cancer initiation is associated with shorter telomere length in Li-Fraumeni syndrome. *Cancer Res* **67,** 1415–1418 (2007).
157. Martinez-Delgado, B. *et al.* Genetic anticipation is associated with telomere shortening in hereditary breast cancer. *PLoS Genet* **7,** e1002182 (2011).
158. Calado, R. T. *et al.* Constitutional hypomorphic telomerase mutations in patients with acute myeloid leukemia. *Proc Natl Acad Sci USA* **106,** 1187–1192 (2009).
159. Kitajima, K. *et al.* Redirecting differentiation of hematopoietic progenitors by a transcription factor, GATA-2. *Blood* **107,** 1857–1863 (2006).
160. Yan, X.-J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat Genet* **43,** 309–315 (2011).
161. Zhang, S.-J. *et al.* Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia. *Proc Natl Acad Sci USA* **105,** 2076–2081 (2008).
162. Ostergaard, P. *et al.* Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nat Genet* (2011). doi:10.1038/ng.923
163. Hsu, A. P. *et al.* Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. *Blood* **118,** 2653–2655 (2011).
164. Dickinson, R. E. *et al.* Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. *Blood* **118,** 2656–2658 (2011).
165. Bödör, C. *et al.* Germ-line GATA2 p.THR354MET mutation in familial myelodysplastic syndrome with acquired monosomy 7 and ASXL1 mutation demonstrating rapid onset and poor survival. *Haematologica* **97,** 890–894 (2012).
166. Kazenwadel, J. *et al.* Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature. *Blood* (2011). doi:10.1182/blood-2011-08-374363
167. Horwitz, M. S. GATA2 deficiency: flesh and blood. *Blood* **123,** 799–800 (2014).
168. West, R. R., Hsu, A. P., Holland, S. M., Cuellar-Rodriguez, J. & Hickstein, D. D. Acquired ASXL1 mutations are common in patients with inherited GATA2 mutations and correlate with myeloid transformation. *Haematologica* **99,** 276–281 (2014).
169. Gonzalez, K. D. *et al.* Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *J. Clin. Oncol.* **27,** 1250–1256 (2009).
170. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
171. Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of

somatic mutations in cancer. *Nat Rev Genet* **14,** 703–718 (2013).
172. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45,** 1127–1133 (2013).
173. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).
174. The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N Engl J Med* (2013). doi:10.1056/NEJMoa1301689
175. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478,** 64–69 (2011).
176. Papaemmanuil, E. *et al.* Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365,** 1384–1395 (2011).
177. Fabbri, G. *et al.* Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation. *J. Exp. Med.* **208,** 1389–1401 (2011).
178. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* (2011). doi:10.1038/nature10113
179. Wang, L. *et al.* SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* **365,** 2497–2506 (2011).
180. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44,** 47–52 (2012).
181. Pasqualucci, L. *et al.* Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* **471,** 189–195 (2011).
182. Pasqualucci, L. *et al.* Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat Genet* **43,** 830–837 (2011).
183. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
184. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci USA* **109,** 3879–3884 (2012).
185. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
186. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nat Rev Genet* **12,** 465–474 (2011).
187. Powell, B. C. *et al.* Identification of TP53 as an acute lymphocytic leukemia susceptibility gene through exome sequencing. *Pediatr. Blood Cancer* **60,** E1–3 (2013).
188. Perez-Garcia, A. *et al.* Genetic loss of SH2B3 in acute lymphoblastic leukemia. *Blood* **122,** 2425–2432 (2013).
189. Shah, S. *et al.* A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat Genet* **45,** 1226–U179 (2013).
190. Kirwan, M. J. *et al.* Exome Sequencing Identifies Autosomal-Dominant SRP72 Mutations Associated with Familial Aplasia and Myelodysplasia. *Am J Hum Genet* **90,** 888–892 (2012).
191. Saarinen, S. *et al.* Exome sequencing reveals germline NPAT mutation as a candidate risk factor for Hodgkin lymphoma. *Blood* **118,** 493–498 (2011).
192. Saarinen, S. *et al.* Primary mediastinal large B-cell lymphoma segregating in a family: exome sequencing identifies MLL as a candidate predisposition gene. *Blood* **121,** 3428–3430 (2013).

193. Auer, F. *et al.* Inherited susceptibility to pre B-ALL caused by germline transmission of PAX5 c.547G>A. *Leukemia* **28,** 1136–1138 (2014).
194. O'Brien, P., Morin, P., Ouellette, R. J. & Robichaud, G. A. The Pax-5 gene: a pluripotent regulator of B-cell differentiation and cancer disease. *Cancer Res* **71,** 7345–7350 (2011).
195. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
196. Nebral, K. *et al.* Incidence and diversity of PAX5 fusion genes in childhood acute lymphoblastic leukemia. *Leukemia* **23,** 134–143 (2009).
197. Reade, C. J., Riva, J. J., Busse, J. W., Goldsmith, C. H. & Elit, L. Risks and benefits of screening asymptomatic women for ovarian cancer: a systematic review and meta-analysis. *Gynecol. Oncol.* **130,** 674–681 (2013).
198. Harding, C. *et al.* Breast Cancer Screening, Incidence, and Mortality Across US Counties. *JAMA Intern Med* (2015). doi:10.1001/jamainternmed.2015.3043
199. Burn, J., Mathers, J. C. & Bishop, D. T. Chemoprevention in Lynch syndrome. *Fam. Cancer* **12,** 707–718 (2013).
200. Australian Bureau of Statistics. Australian Demographic Statistics, Sep 2013. at <http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0/>
201. Australian Bureau of Statistics. Migration, Australia, 2011-12 and 2012-13. at <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/BEF8BD30A177EC39CA257C4400238EED?opendocument>
202. Department of Immigration and Border Protection. The People of Tasmania. 1–298 (2014). doi:978-1-920996-29-1
203. The companion to Tasmanian history. (2006). at <http://www.utas.edu.au/library/companion_to_tasmanian_history/index.htm>
204. Pridmore, S. A. The large Huntington's disease family of Tasmania. *Med. J. Aust.* **153,** 593–595 (1990).
205. Shepherd, J. J. The Natural-History of Multiple Endocrine Neoplasia Type-1 - Highly Uncommon or Highly Unrecognized. *Arch Surg* **126,** 935–952 (1991).
206. FitzGerald, L. M. *et al.* Identification of a prostate cancer susceptibility gene on chromosome 5p13q12 associated with risk of both familial and sporadic disease. *Eur J Hum Genet* **17,** 368–377 (2009).
207. Green, C. M. *et al.* How significant is a family history of glaucoma? Experience from the Glaucoma Inheritance Study in Tasmania. *Clin. Experiment. Ophthalmol.* **35,** 793–799 (2007).
208. Rubio, J. P. *et al.* Genetic dissection of the human leukocyte antigen region by use of haplotypes of Tasmanians with multiple sclerosis. *Am J Hum Genet* **70,** 1125–1137 (2002).
209. Lickiss, J. N., Baikie, A. G. & Panton, J. Lymphoproliferative and myeloproliferative disease in Tasmania. *Natl Cancer Inst Monogr* **47,** 37–39 (1977).
210. Lickiss, J. N. *et al.* Myeloproliferative and lymphoproliferative disorders in Tasmania, 1972-80: patterns in space and time. *J Natl Cancer Inst* **72,** 1223–1231 (1984).
211. Giles, G. G., Lickiss, J. N., Baikie, M. J., Lowenthal, R. M. & Panton, J. Myeloproliferative and lymphoproliferative disorders in Tasmania, 1972-80: occupational and familial aspects. *J Natl Cancer Inst* **72,** 1233–1240 (1984).

212. Lowenthal, R. M., Tegg, E. M. & Dickinson, J. L. The Familial Tasmanian Haematological Malignancies Study (FaTHMS) : Its origins, its history and the phenomenon of anticipation. *Transfus. Apher. Sci.* (2013). doi:10.1016/j.transci.2013.07.011
213. National Health and Medical Research Council, Australian Research CouncilCommittee, Australian Vice-Chancellors. *National Statement on Ethical Conduct in Human Research*. (NHMRC, 2007).
214. National Health and Medical Research Council, Australian Research CouncilAustralia, Universities. *Australian Code for the Responsible Conduct of Research*. *nhmrc.gov.au* (2007).
215. Callisaya, M. L. *et al.* Gait, gait variability and the risk of multiple incident falls in older people: a population-based study. *Age Ageing* **40,** 481–487 (2011).
216. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* **132,** 1077–1130 (2013).
217. Wang, S. R. *et al.* Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. *Am J Hum Genet* **94,** 710–720 (2014).
218. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5,** 28 (2013).
219. Bao, S. *et al.* Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet* **56,** 406–414 (2011).
220. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
221. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297–1303 (2010).
222. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
223. Hu, H. *et al.* A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nat Biotechnol* **32,** 663–669 (2014).
224. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res* **21,** 1529–1542 (2011).
225. Hu, H. *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.* **37,** 622–634 (2013).
226. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164 (2010).
227. Singleton, M. V. *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* **94,** 599–610 (2014).
228. Illumina. Improved Accuracy for ELAND and Variant Calling. *Technical Note Sequencing* 1–8 (2011). at <http://res.illumina.com/documents/products/technotes/technote_eland_variantcalling_improvements.pdf>
229. GNU Make. *gnu.org*

230. Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data. *bioinformatics.bbsrc.ac.uk* (2010). at <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>
231. Garcia-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28,** 2678–2679 (2012).
232. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
233. R Core Team. R: A Language and Environment for Statistical Computing. (2014). at <http://www.R-project.org>
234. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42,** D764–D770 (2014).
235. Kundaje, A. A comprehensive collection of signal artifact blacklist regions in the human genome. (2013). at <http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf>
236. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
237. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42,** D756–63 (2014).
238. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med* **5,** 100 (2013).
239. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
240. Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res* **41,** D56–63 (2013).
241. *Catalogue of Somatic Mutations in Cancer - COSMIC*. (Wellcome Trust Sanger Insitute). at <http://www.sanger.ac.uk/genetics/CGP/cosmic/>
242. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34,** E2393–402 (2013).
243. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29,** 308–311 (2001).
244. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).
245. Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N. & Brudno, M. Identification of deleterious synonymous variants in human genomes. *Bioinformatics* **29,** 1843–1850 (2013).
246. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **38,** D211–22 (2010).
247. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
248. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30,** 523–530 (2014).
249. An, O. *et al.* NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database (Oxford)* **2014,** bau015–bau015 (2014).
250. Venselaar, H., Beek, Te, T. A. H., Kuipers, R. K. P., Hekkelman, M. L. &

Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11,** 548 (2010).
251. Baralle, D. & Baralle, M. Splicing in action: assessing disease causing sequence changes. *Journal of medical genetics* **42,** 737–748 (2005).
252. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* **10,** R130 (2009).
253. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101,** 6062–6067 (2004).
254. Kennedy, B. *et al.* Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Protoc Hum Genet* **81,** 6.14.1–6.14.25 (2014).
255. Park, D. J. *et al.* Rare Mutations in RINT1 Predispose Carriers to Breast and Lynch Syndrome-Spectrum Cancers. *Cancer Discov* **4,** 804–815 (2014).
256. Mansur, M. B. *et al.* Occurrence of identical NOTCH1 mutation in non-twinned sisters with T-cell acute lymphoblastic leukemia. *Leukemia* **25,** 1368–1370 (2011).
257. Bigas, A., D'Altri, T. & Espinosa, L. The Notch pathway in hematopoietic stem cells. *Curr. Top. Microbiol. Immunol.* **360,** 1–18 (2012).
258. Giambra, V. *et al.* NOTCH1 promotes T cell leukemia-initiating activity by RUNX-mediated regulation of PKC-θ and reactive oxygen species. *Nat Med* **18,** 1693–1698 (2012).
259. Mansouri, L. *et al.* NOTCH1 and SF3B1 mutations can be added to the hierarchical prognostic classification in chronic lymphocytic leukemia. *Leukemia* (2012). doi:10.1038/leu.2012.307
260. Del Poeta, G. *et al.* Clinical significance of c.7544-7545 delCT NOTCH1 mutation in chronic lymphocytic leukaemia. *Br J Haematol* (2012). doi:10.1111/bjh.12128
261. Alkindy, A., Chuzhanova, N., Kini, U., Cooper, D. N. & Upadhyaya, M. Genotype-phenotype associations in neurofibromatosis type 1 (NF1): an increased risk of tumor complications in patients with NF1 splice-site mutations? *Hum. Genomics* **6,** 12 (2012).
262. Side, L. E. *et al.* Mutations of the NF1 gene in children with juvenile myelomonocytic leukemia without clinical evidence of neurofibromatosis, type 1. *Blood* **92,** 267–272 (1998).
263. Balgobind, B. V. *et al.* Leukemia-associated NF1 inactivation in patients with pediatric T-ALL and AML lacking evidence for neurofibromatosis. *Blood* **111,** 4322–4328 (2008).
264. Krieg, A., Le Negrate, G. & Reed, J. C. RIP2-beta: a novel alternative mRNA splice variant of the receptor interacting protein kinase RIP2. *Mol. Immunol.* **46,** 1163–1170 (2009).
265. Miyata, Y. *et al.* Met in urological cancers. *Cancers (Basel)* **6,** 2387–2403 (2014).
266. Shen, Q. *et al.* NAT10, a nucleolar protein, localizes to the midbody and regulates cytokinesis and acetylation of microtubules. *Exp. Cell Res.* **315,** 1653–1667 (2009).
267. Shiratori, A. *et al.* Assignment of the 49-kDa (PRIM1) and 58-kDa (PRIM2A and PRIM2B) subunit genes of the human DNA primase to chromosome bands 1q44 and 6p11.1-p12. *Genomics* **28,** 350–353 (1995).
268. Yatsula, B., Galvao, C., McCrann, M. & Perkins, A. S. Assessment of F-

MuLV-induced tumorigenesis reveals new candidate tumor genes including Pecam1, St7, and Prim2. *Leukemia* **20,** 162–165 (2006).
269. Robasky, K., Lewis, N. E. & Church, G. M. The role of replicates for error mitigation in next-generation sequencing. *Nat Rev Genet* **15,** 56–62 (2014).
270. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13,** 134 (2012).
271. NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server. *evs.gs.washington.edu* at <http://evs.gs.washington.edu/EVS/>
272. Cortes-Ledesma, F., Khamisy, El, S. F., Zuma, M. C., Osborn, K. & Caldecott, K. W. A human 5'-tyrosyl DNA phosphodiesterase that repairs topoisomerase-mediated DNA damage. *Nature* **461,** 674–678 (2009).
273. Palma, C., Binaschi, M., Bigioni, M., Maggi, C. A. & Goso, C. CD137 and CD137 ligand constitutively coexpressed on human T and B leukemia cells signal proliferation and survival. *Int. J. Cancer* **108,** 390–398 (2004).
274. *exac.broadinstitute.org*. at <http://exac.broadinstitute.org>
275. Schellenberg, M. J. *et al.* Mechanism of repair of 5 '-topoisomerase II-DNA adducts by mammalian tyrosyl- DNA phosphodiesterase 2. *Nat. Struct. Mol. Biol.* **19,** 1363–+ (2012).
276. Gomez-Herreros, F. *et al.* TDP2-Dependent Non-Homologous End-Joining Protects against Topoisomerase II-Induced DNA Breaks and Genome Instability in Cells and In Vivo. *PLoS Genet* **9,** (2013).
277. Won, E.-Y. *et al.* The Structure of the Trimer of Human 4-1BB Ligand Is Unique among Members of the Tumor Necrosis Factor Superfamily. *J. Biol. Chem.* **285,** 9202–9210 (2010).
278. Schwarz, H., Blanco, F. J., Kempis, von, J., Valbracht, J. & Lotz, M. ILA, a member of the human nerve growth factor/tumor necrosis factor receptor family, regulates T-lymphocyte proliferation and survival. *Blood* **87,** 2839–2845 (1996).
279. Michel, J., Pauly, S., Langstein, J., Krammer, P. H. & Schwarz, H. CD137-induced apoptosis is independent of CD95. *Immunology* **98,** 42–46 (1999).
280. Scholtysik, R. *et al.* Recurrent deletions of the TNFSF7 and TNFSF9 genes in 19p13.3 in diffuse large B-cell and Burkitt lymphomas. *Int. J. Cancer* **131,** E830–5 (2012).
281. Middendorp, S. *et al.* Mice deficient for CD137 ligand are predisposed to develop germinal center-derived B-cell lymphoma. *Blood* **114,** 2280–2289 (2009).
282. Blyth, K., Cameron, E. R. & Neil, J. C. The RUNX genes: gain or loss of function in cancer. *Nat Rev Cancer* **5,** 376–387 (2005).
283. Moyzis, R. K. *et al.* A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc Natl Acad Sci USA* **85,** 6622–6626 (1988).
284. Counter, C. M. *et al.* Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J* **11,** 1921–1929 (1992).
285. Harley, C. B., Futcher, A. B. & Greider, C. W. Telomeres shorten during ageing of human fibroblasts. *Nature* **345,** 458–460 (1990).
286. Greider, C. W. & Blackburn, E. H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell* **43,** 405–413 (1985).
287. Allsopp, R. C. *et al.* Telomere length predicts replicative capacity of human

fibroblasts. *Proc Natl Acad Sci USA* **89,** 10114–10118 (1992).
288. Brouilette, S. W. *et al.* Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study. *Lancet* **369,** 107–114 (2007).
289. Hartmann, U. *et al.* Telomere length and hTERT expression in patients with acute myeloid leukemia correlates with chromosomal abnormalities. *Haematologica* **90,** 307–316 (2005).
290. Wu, K., Lund, M., Bang, K. & Thestrup-Pedersen, K. Telomerase activity and telomere length in lymphocytes from patients with cutaneous T-cell lymphoma. *Cancer* **86,** 1056–1063 (1999).
291. Wu, K.-D. *et al.* Telomerase and telomere length in multiple myeloma: correlations with disease heterogeneity, cytogenetic status, and overall survival. *Blood* **101,** 4982–4989 (2003).
292. Terasaki, Y., Okumura, H., Ohtake, S. & Nakao, S. Accelerated telomere length shortening in granulocytes: a diagnostic marker for myeloproliferative diseases. *Exp. Hematol.* **30,** 1399–1404 (2002).
293. Ghaffari, S. H., Shayan-Asl, N., Jamialahmadi, A. H., Alimoghaddam, K. & Ghavamzadeh, A. Telomerase activity and telomere length in patients with acute promyelocytic leukemia: indicative of proliferative activity, disease progression, and overall survival. *Ann. Oncol.* **19,** 1927–1934 (2008).
294. Capraro, V. *et al.* Telomere deregulations possess cytogenetic, phenotype, and prognostic specificities in acute leukemias. *Exp. Hematol.* **39,** 195–202.e2 (2011).
295. Lan, Q. *et al.* A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of non-Hodgkin lymphoma. *Clin. Cancer Res.* **15,** 7429–7433 (2009).
296. Mansouri, L. *et al.* Short telomere length is associated with NOTCH1/SF3B1/TP53 aberrations and poor outcome in newly diagnosed chronic lymphocytic leukemia patients. *Am. J. Hematol.* **88,** 647–651 (2013).
297. McGrath, M., Wong, J. Y. Y., Michaud, D., Hunter, D. J. & de Vivo, I. Telomere length, cigarette smoking, and bladder cancer risk in men and women. *Cancer Epidemiol Biomarkers Prev* **16,** 815–819 (2007).
298. Willeit, P. *et al.* Telomere length and risk of incident cancer and cancer mortality. *JAMA* **304,** 69–75 (2010).
299. Artandi, S. E. *et al.* Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406,** 641–645 (2000).
300. Wu, X. *et al.* Telomere dysfunction: a potential cancer predisposition factor. *J Natl Cancer Inst* **95,** 1211–1218 (2003).
301. Slagboom, P. E., Droog, S. & Boomsma, D. I. Genetic determination of telomere size in humans: a twin study of three age groups. *Am J Hum Genet* **55,** 876–882 (1994).
302. Vasa-Nicotera, M. *et al.* Mapping of a major locus that determines telomere length in humans. *Am J Hum Genet* **76,** 147–151 (2005).
303. Njajou, O. T. *et al.* Telomere length is paternally inherited and is associated with parental lifespan. *Proc Natl Acad Sci USA* **104,** 12135–12139 (2007).
304. Huda, N., Tanaka, H., Herbert, B.-S., Reed, T. & Gilley, D. Shared environmental factors associated with telomere length maintenance in elderly male twins. *Aging Cell* **6,** 709–713 (2007).
305. Graakjaer, J. *et al.* The pattern of chromosome-specific variations in telomere length in humans is determined by inherited, telomere-near factors and is

maintained throughout life. *Mech Ageing Dev* **124,** 629–640 (2003).
306. Graakjaer, J., Londoño-Vallejo, J. A., Christensen, K. & Kølvraa, S. The pattern of chromosome-specific variations in telomere length in humans shows signs of heritability and is maintained through life. *Ann N Y Acad Sci* **1067,** 311–316 (2006).
307. Chiang, Y. J. *et al.* Telomere length is inherited with resetting of the telomere set-point. *Proc Natl Acad Sci USA* **107,** 10148–10153 (2010).
308. Aubert, G., Hills, M. & Lansdorp, P. M. Telomere length measurement-Caveats and a critical assessment of the available technologies and tools. *Mutat Res* (2011). doi:10.1016/j.mrfmmm.2011.04.003
309. Kimura, M. *et al.* Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat Protoc* **5,** 1596–1607 (2010).
310. Cawthon, R. M. Telomere measurement by quantitative PCR. *Nucleic Acids Res* **30,** e47 (2002).
311. Cawthon, R. M. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res* **37,** e21 (2009).
312. Almasy, L., Almasy, L., Blangero, J. & Blangero, J. Variance component methods for analysis of complex phenotypes. *Cold Spring Harb Protoc* **2010,** pdb.top77–pdb.top77 (2010).
313. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62,** 1198–1211 (1998).
314. Ruijter, J. M. *et al.* Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res* **37,** e45 (2009).
315. Njajou, O. T. *et al.* A common variant in the telomerase RNA component is associated with short telomere length. *PLoS ONE* **5,** e13048 (2010).
316. Kampstra, P. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* **28,** (2008).
317. Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Res* **38,** D492–6 (2010).
318. Mirabello, L. *et al.* The Association of Telomere Length and Genetic Variation in Telomere Biology Genes. *Hum Mutat* **31,** 1050–1058 (2010).
319. Aubert, G., Baerlocher, G. M., Vulto, I., Poon, S. S. & Lansdorp, P. M. Collapse of Telomere Homeostasis in Hematopoietic Cells Caused by Heterozygous Mutations in Telomerase Genes. *PLoS Genet* **8,** e1002696 (2012).
320. Schröder, C. P. *et al.* Telomere length in breast cancer patients before and after chemotherapy with or without stem cell transplantation. *Br J Cancer* **84,** 1348–1353 (2001).
321. Mirabello, L. *et al.* Leukocyte telomere length in a population-based case-control study of ovarian cancer: a pilot study. *Cancer Causes Control* **21,** 77–82 (2010).
322. Engelhardt, M. *et al.* Telomerase activity and telomere length in pediatric patients with malignancies undergoing chemotherapy. *Leukemia* **12,** 13–24 (1998).
323. Franco, S. *et al.* Telomere dynamics in childhood leukemia and solid tumors: a follow-up study. *Leukemia* **17,** 401–410 (2003).
324. Trkova, M., Prochazkova, K., Krutilkova, V., Sumerauer, D. & Sedlacek, Z. Telomere length in peripheral blood cells of germline TP53 mutation carriers is shorter than that of normal individuals of corresponding age. *Cancer* **110,**

694–702 (2007).
325. Diker-Cohen, T. *et al.* The effect of chemotherapy on telomere dynamics: clinical results and possible mechanisms. *Leuk Lymphoma* (2013). doi:10.3109/10428194.2012.757765
326. Codd, V. *et al.* Common variants near TERC are associated with mean telomere length. *Nat Genet* **42,** 197–199 (2010).
327. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* **45,** 422–7– 427e1–2 (2013).
328. Gu, J. *et al.* A Genome-Wide Association Study Identifies a Locus on Chromosome 14q21 as a Predictor of Leukocyte Telomere Length and as a Marker of Susceptibility for Bladder Cancer. *Cancer Prev Res (Phila)* **4,** 514–521 (2011).
329. Lee, J. H. *et al.* Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: the Long Life Family Study. *Front Genet* **4,** 310 (2013).
330. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc Natl Acad Sci USA* **107,** 9293–9298 (2010).
331. Liu, Y. *et al.* A genome-wide association study identifies a locus on TERT for mean telomere length in Han Chinese. *PLoS ONE* **9,** e85043 (2014).
332. Mangino, M. *et al.* A genome-wide association study identifies a novel locus on chromosome 18q12.2 influencing white cell telomere length. *Journal of medical genetics* **46,** 451–454 (2009).
333. Mangino, M. *et al.* Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. *Hum Mol Genet* **21,** 5385–5394 (2012).
334. Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk. *Hum Mol Genet* (2013). doi:10.1093/hmg/ddt355
335. Prescott, J. *et al.* Genome-wide association study of relative telomere length. *PLoS ONE* **6,** e19635 (2011).
336. Saxena, R. *et al.* Genome-wide association study identifies variants in casein kinase II (CSNK2A2) to be associated with leukocyte telomere length in a Punjabi Sikh diabetic cohort. *Circ Cardiovasc Genet* **7,** 287–295 (2014).
337. Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science* (2013). doi:10.1126/science.1230062
338. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4,** 2185 (2013).
339. Armanios, M. Syndromes of telomere shortening. *Annu Rev Genomics Hum Genet* **10,** 45–61 (2009).
340. Armanios, M. & Blackburn, E. H. The telomere syndromes. *Nat Rev Genet* **13,** 693–704 (2012).
341. Bozzao, C. *et al.* Analysis of telomere dynamics in peripheral blood cells from patients with Lynch syndrome. *Cancer* **117,** 4325–4335 (2011).
342. Rumi, E. *et al.* Disease anticipation in familial myeloproliferative neoplasms. *Blood* **112,** 2587–2588 (2008).
343. Goyama, S., Huang, G., Kurokawa, M. & Mulloy, J. C. Posttranslational modifications of RUNX1 as potential anticancer targets. *Oncogene* **0,** (2014).
344. The Australasian Leukaemia & Lymphoma Group (ALLG). at

<http://www.allg.org.au>
345. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2,** 401–404 (2012).
346. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6,** pl1 (2013).
347. Zhao, S., Xing, Y. & Natkunam, Y. Use of CD137 ligand expression in the detection of small B-cell lymphomas involving the bone marrow. *Hum. Pathol.* **45,** 1024–1030 (2014).
348. Cheng, K. *et al.* CD137 ligand signalling induces differentiation of primary acute myeloid leukaemia cells. *Br J Haematol* **165,** 134–144 (2014).
349. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Publishing Group* **7,** 248–249 (2010).
350. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4,** 1073–1081 (2009).
351. Xie, F. *et al.* Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Res* **24,** gr.173427.114–1533 (2014).
352. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
353. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
354. Migliorini, G. *et al.* Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood* **122,** 3298–3307 (2013).
355. Yang, J. J. *et al.* Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. *Blood* **120,** 4197–4204 (2012).
356. Treviño, L. R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41,** 1001–1005 (2009).
357. Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet* **41,** 1006–1010 (2009).
358. Knight, J. A. *et al.* Genome-wide association study to identify novel loci associated with therapy-related myeloid leukemia susceptibility. *Blood* **113,** 5575–5582 (2009).
359. Tan, D. E. K. *et al.* Genome-wide association study of B cell non-Hodgkin lymphoma identifies 3q27 as a susceptibility locus in the Chinese population. *Nat Genet* **45,** 804–807 (2013).
360. Crowther-Swanepoel, D. *et al.* Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42,** 132–136 (2010).
361. Slager, S. L. *et al.* Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia. *Blood* **120,** 843–846 (2012).
362. Slager, S. L. *et al.* Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood* **117,** 1911–1916 (2011).
363. Kim, D. H. *et al.* A genome-wide association study identifies novel loci associated with susceptibility to chronic myeloid leukemia. *Blood* (2011).

doi:10.1182/blood-2011-01-329797
364. Skibola, C. F. *et al.* Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat Genet* **41,** 873–875 (2009).
365. Conde, L. *et al.* Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet* **42,** 661–664 (2010).
366. Smedby, K. E. *et al.* GWAS of Follicular Lymphoma Reveals Allelic Heterogeneity at 6p21.32 and Suggests Shared Genetic Susceptibility with Diffuse Large B-cell Lymphoma. *PLoS Genet* **7,** e1001378 (2011).
367. Enciso-Mora, V. *et al.* A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). *Nat Genet* **42,** 1126–1130 (2010).
368. Urayama, K. Y. *et al.* Genome-wide association study of classical Hodgkin lymphoma and Epstein-Barr virus status-defined subgroups. *J Natl Cancer Inst* **104,** 240–253 (2012).
369. Frampton, M. *et al.* Variation at 3p24.1 and 6q23.3 influences the risk of Hodgkin's lymphoma. *Nat Commun* **4,** 2549 (2013).
370. Kumar, V. *et al.* Common variants on 14q32 and 13q12 are associated with DLBCL susceptibility. *J Hum Genet* (2011). doi:10.1038/jhg.2011.35
371. Vijai, J. *et al.* Susceptibility Loci associated with specific and shared subtypes of lymphoid malignancies. *PLoS Genet* **9,** e1003220 (2013).
372. Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat Genet* **44,** 58–61 (2012).
373. Chubb, D. *et al.* Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat Genet* **45,** 1221–1225 (2013).
374. Weinhold, N. *et al.* The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat Genet* **45,** 522–525 (2013).
375. Kilpivaara, O. *et al.* A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nat Genet* **41,** 455–459 (2009).
376. Cozen, W. *et al.* A genome-wide meta-analysis of nodular sclerosing Hodgkin lymphoma identifies risk loci at 6p21.32. *Blood* **119,** 469–475 (2012).
377. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).

Appendices

Appendix 1.1 Summary of the GWAS significant loci from HM studies

HM Subtype	Initial Sample Size	Replication Sample Size	Region	hg19 Chromosome and position	RefSeq gene location or genes within +/- 10 kb when intergenic	RefSeq gene region	SNP rsIDs	p-Value	p-Value note	OR [95% CI] (NR = not reported)	Platform [SNPs passing QC]	Reference
ALL	1,658 European ancestry child cases, 4,723 European ancestry controls	1,449 European ancestry child cases, 1,488 European ancestry controls	10p12.2	chr10:22839628	*PIP4K2A*	exonic	rs2230469	2×10^{-9}		1.23 [1.15-1.32]	Illumina [382,776] (imputed)	354
			10q21.2	chr10:63721176	*ARID5B*	downstream	rs7090445	5×10^{-54}		NR		
			10p14	chr10:8104208	*GATA3*	intronic	rs3824662	9×10^{-12}		1.31 [1.21-1.41]		
			14q11.2	chr14:23589057	*CEBPE*	intronic	rs2239633	1×10^{-16}		NR		
			7p12.2	chr7:50469981	*IKZF1*	intronic	rs11980379	3×10^{-33}		NR		
			9p21.3	chr9:21984661	*CDKN2A*	intronic	rs3731217	2×10^{-8}		NR		
ALL	1,268 cases of European, South Asian, East Asian, African American/Afro-Caribbean, Hispanic and other ancestries	1,267 cases of European, South Asian, East Asian, African American/Afro-Caribbean, Hispanic and other ancestries	1p31.3	chr1:66762466	*PDE4B*	intronic	rs546784	9×10^{-6}		1.40 [1.20-1.62]	Affymetrix [444,044]	355
			1p31.3	chr1:66769100	*PDE4B*	intronic	rs6683977	5×10^{-6}		1.41 [1.22-1.64]		
			10q22.1	chr10:72040805	*NPFFR1*	intronic	rs41322152	8×10^{-6}		2.52 [1.68-3.79]		
			14q22.1	chr14:51403531	*PYGL*	intronic	rs7142143	7×10^{-9}		3.61 [2.34-5.57]		
			18q12.3	chr18:40591084	*RIT2*	intronic	rs9958208	5×10^{-6}		1.62 [1.32-1.99]		
			2q23.2	chr2:150397218	intergenic	intergenic	rs7578361	8×10^{-6}		1.40 [1.21-1.63]		
			2q23.2	chr2:150457624	*LOC101929231*	intronic	rs10170236	4×10^{-6}		1.45 [1.24-1.69]		
			3p22.1	chr3:39847072	*MYRIP*	upstream	rs17079534	2×10^{-7}		4.07 [2.40-6.87]		
			7q34	chr7:139702593	*TBXAS1*	intronic	rs17837497	2×10^{-6}		2.34 [1.65-3.31]		
			7q21.11	chr7:78331465	*MAGI2*	intronic	rs1496766	5×10^{-6}		2.84 [1.81-4.44]		
ALL	317 European descent cases, 17,958 European descent controls	NA	1q31.3	chr1:196844593	*CFHR4*	intronic	rs6428370	7×10^{-6}		1.43 [1.20-1.60]	Affymetrix [307,944]	356
			1q43	chr1:237266603	*RYR2*	intronic	rs7554607	2×10^{-6}		1.49 [1.20-1.70]		
			1q44	chr1:247689532	*GCSAML, GCSAML-AS1*	intronic	rs1881797	7×10^{-6}		1.52 [1.20-1.80]		
			1p31.1	chr1:76772328	*ST6GALNAC3*	intronic	rs10873876	4×10^{-6}		1.55 [1.20-1.80]		
			10p11.21	chr10:34817988	*PARD3*	intronic	rs563507	9×10^{-6}		2.00 [1.40-2.70]		
			10q21.2	chr10:63723577	*ARID5B*	intronic	rs10821936	1×10^{-15}		1.91 [1.60-2.20]		
			12q24.22	chr12:117002658	*MAP1LC3B2*	intronic	rs2089222	8×10^{-8}		2.26 [1.60-3.00]		
			12p13.32	chr12:4425122	*c12orf5*	upstream	rs10849033	9×10^{-6}		2.55 [1.60-3.80]		
			18p11.32	chr18:2498054	intergenic	intergenic	rs1879352	9×10^{-6}		1.53 [1.20-1.80]		
			19q13.31	chr19:44511389	*ZNF230*	intronic	rs2191566	4×10^{-7}		1.52 [1.20-1.70]		
			2q36.1	chr2:223917983	*KCNE4*	downstream	rs12621643	3×10^{-6}		1.48 [1.20-1.70]		
			3q26.32	chr3:178429939	*KCNMB2*	intronic	rs9290663	6×10^{-6}		1.58 [1.20-1.90]		
			6q24.1	chr6:141169825	intergenic	intergenic	rs11155133	3×10^{-7}		3.62 [2.10-6.00]		
			7p12.2	chr7:50466304	*IKZF1*	intronic	rs11978267	8×10^{-11}		1.69 [1.40-1.90]		
ALL	419 European ancestry cases, 474 European ancestry controls	Up to 1,277 European ancestry cases and 3,061 European ancestry controls	11p11.2	chr11:48115089	*PTPRJ*	intronic	rs3942852	1×10^{-6}	*ETV6-RUNX1* positive	NR	Affymetrix [355,750]	21
			11q12.1	chr11:56175671	*OR5R1*	downstream	rs1945213	3×10^{-8}	Combined ALL	1.30 [1.19-1.43]		
			11q12.1	chr11:56175671	*OR5R1*	downstream	rs1945213	4×10^{-8}	*ETV6-RUNX1* positive	NR		
			14q24.3	chr14:76703351	intergenic	intergenic	rs7156960	3×10^{-6}	Combined ALL	1.22 [1.12-1.33]		
			14q24.3	chr14:76703351	intergenic	intergenic	rs7156960	1×10^{-6}	*ETV6-RUNX1* positive	NR		
			15q26.1	chr15:92657373	*SLCO3A1*	intronic	rs207954	1×10^{-6}	*ETV6-RUNX1* positive	NR		
			3q28	chr3:189401776	*TP63*	3'UTR	rs17505102	2×10^{-8}	Combined ALL	1.47 [1.28-1.67]		
			3q28	chr3:189401776	*TP63*	3'UTR	rs17505102	9×10^{-9}	*ETV6-RUNX1* positive	NR		
			4q13.1	chr4:59503726	intergenic	intergenic	rs282708	8×10^{-6}	Combined ALL	1.23 [1.12-1.35]		
			6q14.1	chr6:77789808	intergenic	intergenic	rs7738636	6×10^{-6}	Combined ALL	1.27 [1.15-1.41]		
			8p21.3	chr8:19651161	intergenic	intergenic	rs920590	2×10^{-6}	*ETV6-RUNX1* positive	NR		
ALL	907 European ancestry cases, 2,398 European ancestry controls	NA	10q21.2	chr10:63752159	*ARID5B*	intronic	rs7089424	7×10^{-19}		1.65 [1.54-1.76]	Illumina [291,473]	357
			14q11.2	chr14:23589057	*CEBPE*	intronic	rs2239633	3×10^{-7}		1.34 [1.22-1.45]		
			7p12.2	chr7:50470604	*IKZF1*	intronic	rs4132601	1×10^{-19}		1.69 [1.58-1.81]		

HM Subtype	Initial Sample Size	Replication Sample Size	Region	hg19 Chromosome and position	RefSeq gene location or genes within +/- 10 kb when intergenic	RefSeq gene region	SNP rsIDs	p-Value	p-Value note	OR [95% CI] (NR = not reported)	Platform [SNPs passing QC]	Reference
ALL	972 European ancestry cases, 1,386 European ancestry controls, 89 African American cases, 1,363 African American controls, 305 Hispanic cases, 1,008 Hispanic controls	574 European ancestry cases, 2,601 European ancestry controls, 128 African American cases, 1,075 African American controls, 143 Hispanic cases, 640 Hispanic controls	10q21.2	chr10:63723577	*ARID5B*	intronic	rs10821936	6×10^{-46}		1.86 [1.71-2.03]	Affymetrix [709,059]	139
			14q11.2	chr14:23585333	*CEBPE*	intronic	rs4982731	9×10^{-12}		1.36 [1.24-1.48]		
			7p12.2	chr7:50473251	*IKZF1*	3'UTR	rs6964969	2×10^{-29}		1.67 [1.53-1.83]		
AML (therapy-related)	80 European ancestry cases, 150 European ancestry controls	70 European ancestry cases, 95 European ancestry controls	10q21.1	chr10:59949736	*IPMK*	intronic	rs1199098	3.43×10^{-3}		0.46 [0.27-0.79]	Affymetrix [10,136]	358
			17q12	chr17:31789043	*ASIC2*	intronic	rs1394384	4.55×10^{-5}		0.29 [0.15-0.56]		
			3p24.1	chr3:28749314	*LINC00693*	intronic	rs1381392	4.19×10^{-3}		2.08 [1.29-3.35]		
B-cell NHL	253 Singaporean Chinese ancestry cases; 1,438 Singaporean Chinese ancestry controls	1,175 Han Chinese ancestry cases, 5,492 Han Chinese ancestry controls	3q27.3	chr3:187649419	intergenic	intergenic	rs6773854	3×10^{-13}		1.44 [1.31-1.59]	Illumina [550,946]	359
CLL	1,739 European ancestry cases, 5,199 European ancestry controls	1,144 European ancestry cases, 3,151 European ancestry controls	10q23.31	chr10:90749963	*ACTA2*	intronic	rs1800682	2×10^{-8}		1.25 [NR]	Illumina [450,000] (Imputed)	138
			11q24.1	chr11:123361397	intergenic	intergenic	rs735665	4×10^{-24}		1.64 [NR]		
			11p15.5	chr11:2322829	*C11orf21*	intronic	rs11022157	3×10^{-6}		1.22 [NR]		
			15q15.1	chr15:40414116	intergenic	intergenic	rs8023845	2×10^{-6}		1.27 [NR]		
			15q21.3	chr15:56382295	*RFX7*	intronic	rs16976734	4×10^{-7}		1.36 [NR]		
			15q23	chr15:70018990	intergenic	intergenic	rs7176508	8×10^{-18}		1.42 [NR]		
			16q24.1	chr16:85955671	*IRF8*	3'UTR	rs1044873	1×10^{-9}		1.29 [NR]		
			2q13	chr2:111871897	*ACOXL*	intronic	rs1439287	5×10^{-15}		1.37 [NR]		
			2q37.1	chr2:231091223	*SP140*	intronic	rs13397985	5×10^{-13}		1.43 [NR]		
			2p22.2	chr2:37467264	*NDUFAF7*	downstream	rs2041840	9×10^{-6}		1.20 [NR]		
			3q26.2	chr3:169492101	*MYNN*	coding	rs10936599	2×10^{-9}		1.26 [1.17-1.35]		
			4q26	chr4:114683844	*CAMK2D*	3'UTR	rs6858698	3×10^{-9}		1.31 [1.20-1.44]		
			5p15.33	chr5:1344458	*CLPTM1L*	coding	rs31490	2×10^{-7}		1.18 [1.11-1.26]		
			6q25.2	chr6:154478440	*IPCEF1* / *OPRM1*	3'UTR / intronic	rs2236256	2×10^{-10}		1.23 [1.15-1.30]		
			6p21.32	chr6:32257566	intergenic	intergenic	rs926070	4×10^{-8}		1.27 [NR]		
			6p21.31	chr6:33540209	*BAK1*	coding	rs210134	3×10^{-6}		1.31 [NR]		
			6p25.3	chr6:411064	*IRF4*	intronic	rs872071	3×10^{-16}		1.39 [NR]		
			7q31.33	chr7:124462661	*POT1*	downstream	rs17246404	3×10^{-8}		1.22 [1.14-1.31]		
			8q24.21	chr8:128188019	intergenic	intergenic	rs2466024	3×10^{-6}		1.21 [NR]		
CLL	505 European ancestry cases, 1,438 European ancestry controls	1,024 European ancestry cases, 1,677 European ancestry controls	11q24.1	chr11:123361397	intergenic	intergenic	rs735665	4×10^{-12}		1.45 [1.31-1.61]	Illumina [345,665]	137
			15q23	chr15:70018990	intergenic	intergenic	rs7176508	5×10^{-12}		1.37 [1.26-1.50]		
			19q13.32	chr19:47207654	*PRKD2*	intronic	rs11083846	4×10^{-9}		1.35 [1.22-1.49]		
			2q13	chr2:111797458	*ACOXL*	intronic	rs17483466	2×10^{-10}		1.39 [1.25-1.53]		
			2q37.1	chr2:231091223	*SP140*	intronic	rs13397985	6×10^{-10}		1.41 [1.26-1.57]		
			6p25.3	chr6:411064	*IRF4*	intronic	rs872071	2×10^{-20}		1.54 [1.41- 1.69]		
CLL	4 - stage followup of previous study {DiBernardo:2008is}	2503 total European ancestry cases, 5789 total European ancestry controls	15q21.3	chr15:56340896	intergenic	intergenic	rs7169431	4.74×10^{-7}		1.36 [1.21-1.53]	4-stage follow up of previous CLL findings in [137]	360
			16q24.1	chr16:85975659	intergenic	intergenic	rs305061	3.60×10^{-7}		1.22 [1.13-1.31]		
			2q37.3	chr2:242371101	*FARP2*	upstream	rs757978	2.11×10^{-9}		1.39 [1.25-1.56]		
			8q24.21	chr8:128192981	*CASC19*	downstream	rs2456449	7.84×10^{-10}		1.26 [1.17-1.35]		

HM Subtype	Initial Sample Size	Replication Sample Size	Region	hg19 Chromosome and position	RefSeq gene location or genes within +/- 10 kb when intergenic	RefSeq gene region	SNP rsIDs	p-Value	p-Value note	OR [95% CI] (NR = not reported)	Platform [SNPs passing QC]	Reference
CLL	1,121 European ancestry cases, 3,745 European ancestry controls	861 European ancestry cases, 2,033 European ancestry controls	11q24.1	chr11:123361397	intergenic	intergenic	rs735665	3×10^{-12}		1.52 [1.35-1.72]	Affymetrix & Illumina [~1,500,000] (imputed)	361
			15q23	chr15:70018990	intergenic	intergenic	rs7176508	3×10^{-11}		1.42 [1.28-1.58]		
			16q24.1	chr16:85975659	intergenic	intergenic	rs305061	9×10^{-8}		1.33 [1.2-1.49]		
			2q13	chr2:111797458	*ACOXL*	intronic	rs17483466	5×10^{-9}		1.43 [1.27-1.61]		
			2q37.1	chr2:231091223	*SP140*	intronic	rs13397985	2×10^{-7}		1.39 [1.23-1.59]		
			2q37.3	chr2:242371101	*FARP2*	upstream	rs757978	3×10^{-6}		1.46 [1.25-1.72]		
			6p21.31	chr6:33546837	*BAK1*	intronic	rs210142	9×10^{-16}		1.40 [1.25-1.57]		
			6p25.3	chr6:411064	*IRF4*	intronic	rs872071	8×10^{-14}		1.47 [1.33-1.63]		
CLL	407 European ancestry cases, 296 European ancestry controls	252 European ancestry cases, 965 European ancestry controls	16q24.1	chr16:85944439	*IRF8*	intronic	rs391525	3×10^{-9}		1.56 [1.35-1.82]	Affymetrix [827,777]	362
			6p21.32	chr6:32578082	intergenic	intergenic	rs674313	7×10^{-9}		1.69 [1.41-2.01]		
			6p21.32	chr6:32606756	*HLA-DQA1*	intronic	rs9272535	9×10^{-8}		1.61 [1.35-1.92]		
			6p25.3	chr6:417727	*IRF4*	downstream	rs9378805	2×10^{-6}		1.38 [1.20-1.58]		
CLL	2,179 European ancestry cases, 6,221 European ancestry controls	1,709 European ancestry cases, 6,318 European ancestry controls	10q23.31	chr10:90759724	*FAS*	intronic	rs4406737	1×10^{-14}		1.27 [1.19-1.33]	Illumina [549,934]	145
			11q24.1	chr11:123361397	intergenic	intergenic	rs735665	4×10^{-39}		1.62 [NR]		
			11p15.5	chr11:2311152	*C11orf21*	downstream	rs7944004	2×10^{-10}		1.20 [1.13-1.27]		
			15q21.3[	chr15:56775597	intergenic	intergenic	rs11636802	2×10^{-13}		1.41 [NR]		
			15q23	chr15:70018990	intergenic	intergenic	rs7176508	1×10^{-17}		1.32 [NR]		
			16q24.1	chr16:85927814, chr16:85944823, chr16:85975659	*IRF8*	upstream, intronic, downstream	rs391023, rs2292982, rs305061	5×10^{-17}		1.33 [NR]		
			18q21.32	chr18:57622287	intergenic	intergenic	rs4368253	3×10^{-8}		1.19 [1.12-1.27]		
			18q21.33	chr18:60793549	*BCL2*	3'UTR	rs4987855	3×10^{-12}		1.47 [1.32-1.61]		
			18q21.33	chr18:60793921	*BCL2*	intronic	rs4987852	8×10^{-11}		1.41 [NR]		
			2q13	chr2:111616104	*ACOXL*	upstream	rs13401811	2×10^{-18}		1.41 [1.30-1.52]		
			2q13	chr2:111797458	*ACOXL*	intronic	rs17483466	4×10^{-17}		1.37 [NR]		
			2q33.1	chr2:202111380	*CASP8*	intronic	rs3769825	3×10^{-9}		1.19 [1.12-1.25]		
			2q37.1	chr2:231091223	*SP140*	intronic	rs13397985	1×10^{-22}		1.45 [NR]		
			2q37.3	chr2:242371101	*FARP2*	upstream	rs757978	1×10^{-7}		1.29 [NR]		
			4q25	chr4:109016824	*LEF1*	intronic	rs898518	4×10^{-10}		1.20 [1.14-1.27]		
			6p21.32	chr6:32626272, chr6:32578082, chr6:32611641	*HLA-DQA1 / HLA-DQB1*	intergenic, intergenic, 3'UTR	rs9273363, rs674313, rs9273012	2×10^{-10}		1.24 [NR]		
			6p21.31	chr6:33546837	*BAK1*	intronic	rs210142	5×10^{-8}		1.22 [NR]		
			6p25.3	chr6:411064	*IRF4*	intronic	rs872071	6×10^{-20}		1.33 [NR]		
			8q22.3	chr8:103578874	*ODF1*	downstream	rs2511714	5×10^{-8}		1.19 [1.12-1.27]		
			8q24.21	chr8:128211229, chr8:128192981, chr8:128209820	*CASC19*	upstream, downstream, 5'UTR	rs2466035, rs2456449, rs2466032	2×10^{-8}		1.21 [NR]		
			9p21.3	chr9:22206987	intergenic	intergenic	rs1679013	1×10^{-8}		1.19 [1.12-1.27]		
CML	201 Korean ancestry cases, 497 Korean ancestry controls	237 Korean ancestry cases, 1000 Korean ancestry controls, 232 European ancestry cases, 576 European ancestry controls	17q11.1	chr17:25541278	intergenic	intergenic	rs4795519	1×10^{-12}		1.85 [1.56-2.17]	Affymetrix [456,522]	363
			6q25.1	chr6:151907748	*CCDC170*	intronic	rs4869742	2×10^{-6}		1.67 [1.35-2.04]		
FL	189 European ancestry cases, 592 European controls	456 European ancestry cases, 2,785 European ancestry	6p21.33	chr6:31074030	*C6orf15*	downstream	rs6457327	5×10^{-11}		1.69 [1.43-2.00]	Illumina [~500,000] (pooled)	364

HM Subtype	Initial Sample Size	Replication Sample Size	Region	hg19 Chromosome and position	RefSeq gene location or genes within +/- 10 kb when intergenic	RefSeq gene region	SNP rsIDs	p-Value	p-Value note	OR [95% CI] (NR = not reported)	Platform [SNPs passing QC]	Reference
		controls										
FL	681 European ancestry cases, 750 European ancestry controls	up to 3,164 European ancestry cases, 6,208 European ancestry controls	11q24.1	chr11:123361397	intergenic	intergenic	rs735665	4×10^{-9}	CLL/SLL	1.81 [1.50-2.20]	Illumina [312,768]	365
			6p21.33	chr6:31074030	*C6orf15*	downstream	rs6457327	7×10^{-6}	FL	1.47 [1.27-1.72]		
			6p21.32	chr6:32665420	intergenic	intergenic	rs10484561	1×10^{-29}	FL	1.95 [1.72-2.22]		
FL	379 European ancestry cases, 791 European ancestry controls	1,049 European ancestry cases, 3,952 European ancestry controls	6p21.32	chr6:32664458	intergenic	intergenic	rs2647012	2×10^{-21}		1.56 [1.43-1.72]	Illumina [298,168]	366
HL	589 European ancestry cases, 5,199 European ancestry controls	2,057 European ancestry cases, 3,416 European ancestry controls	10p14	chr10:8093034	*GATA3-AS1*	intronic	rs501764	7×10^{-8}		1.25 [1.15-1.36]	Illumina [504,374]	367
			2p16.1	chr2:61066666	*LINC01185*	downstream	rs1432295	2×10^{-8}		1.22 [1.14-1.30]		
			6p21.32	chr6:32428285	intergenic	intergenic	rs6903608	3×10^{-50}		1.70 [1.58-1.82]		
			8q24.21	chr8:129075832	*PVT1*	upstream	rs2608053	1×10^{-7}		1.20 [1.12-1.28]		
			8q24.21	chr8:129192271	intergenic	intergenic	rs2019960	1×10^{-13}		1.33 [1.23-1.44]		
HL	1,200 European ancestry cases, 6,417 European ancestry controls	563 European ancestry cases, 613 European ancestry controls	5q31.1	chr5:131995964	*IL13*	intronic	rs20541	1×10^{-8}	EBV-negative cHL	1.47 [1.29-1.68]	Illumina [502,514]	368
			5q15	chr5:96101944	*CAST / ERAP1*	coding / upstream	rs27524	7×10^{-6}	Total cHL	1.22 [1.11-1.33]		
			6p21.33	chr6:31446796	*HCG26*	downstream	rs2248462	7×10^{-16}	Total cHL	1.64 [1.45-1.85]		
			6p21.32	chr6:32433167	intergenic	intergenic	rs2395185	4×10^{-31}	Total cHL	1.82 [1.67-2.00]		
HL	1,465 European ancestry cases, 6,417 European ancestry controls	1,071 European ancestry cases, 953 cases, 1,853 controls	10p14	chr10:8093034	*GATA3-AS1*	intronic	rs501764	4×10^{-10}		1.39 [NR]	Illumina [296,129]	369
			2p16.1	chr2:61066666	*LINC01185*	downstream	rs1432295	1×10^{-6}		1.24 [NR]		
			3p24.1	chr3:27764623	*EOMES*	intronic	rs3806624	1×10^{-12}		1.26 [1.18-1.34]		
			6q23.3	chr6:135415004	intergenic	intergenic	rs7745098	3×10^{-9}		1.21 [1.14-1.29]		
			6p21.32	chr6:32428285	intergenic	intergenic	rs6903608	5×10^{-27}		1.62 [NR]		
			8q24.21	chr8:129192271	intergenic	intergenic	rs2019960	6×10^{-10}		1.37 [NR]		
DLBCL	74 Japanese ancestry cases, 934 Japanese ancestry controls	325 Japanese ancestry cases, 3,309 Japanese ancestry controls	13q12.2	chr13:28197436	*POLR1D*	intronic	rs7097	7×10^{-6}		1.44 [1.23-1.67]	Illumina [444,361]	370
			14q32.32	chr14:103484825	*CDC42BPB*	intronic	rs751837	3×10^{-7}		3.51 [2.13-5.88]		
Lymphoma	275 FL, 269 DLBCL cases, 198 other NHL cases, 202 HL cases, 4,044 controls	202 European ancestry FL cases, 367 European ancestry DLBCL cases, 577 European ancestry other NHL cases, 99 European ancestry HL cases, 2,596 European ancestry controls	11q12.1	chr11:58060192	intergenic	intergenic	rs12289961	4×10^{-8}	Multiple lymphomas	1.29 [1.17-1.40]	Affymetrix [530,583]	371
			11q12.1	chr11:58060192	intergenic	intergenic	rs12289961	1×10^{-7}	NHL	1.29 [1.17-1.42]		
			11q12.1	chr11:58347765	*ZFP91, ZFP91-CNTF*	intronic	rs948562	6×10^{-7}	Multiple lymphomas	1.29 [1.16-1.43]		
			11q12.1	chr11:58347765	*ZFP91, ZFP91-CNTF*	intronic	rs948562	3×10^{-7}	NHL	1.32 [1.18-1.46]		
			6p23	chr6:14636963	intergenic	intergenic	rs707824	6×10^{-7}	NHL	1.33 [1.17-1.47]		
			6p21.32	chr6:32429643	*C6orf10*	downstream	rs9268853	2×10^{-10}	FL	1.56 [NR]		
			6p21.32	chr6:32581889	intergenic	intergenic	rs4530903	3×10^{-6}	Multiple lymphomas	1.29 [1.16-1.43]		
			6p21.32	chr6:32581889	intergenic	intergenic	rs4530903	2×10^{-8}	NHL	1.37 [1.23-1.54]		
			6p21.32	chr6:32581889	intergenic	intergenic	rs4530903	3×10^{-12}	FL	1.93 [NR]		
			6p21.32	chr6:32668100	intergenic	intergenic	rs2647045	4×10^{-10}	FL	1.69 [NR]		
			6p21.32	chr6:32668336	intergenic	intergenic	rs2647046	2×10^{-6}	NHL	1.25 [1.14-1.37]		
			6p21.32	chr6:32730012	*HLA-DQB2*	intronic	rs7453920	5×10^{-6}	Multiple lymphomas	1.19 [1.11-1.3]		
			6p21.32	chr6:32741868	intergenic	intergenic	rs2621416	2×10^{-9}	FL	1.57 [NR]		
MM	1,675 European ancestry cases, 5,903 European ancestry controls	169 European ancestry cases, 927 European ancestry controls	2p23.3	chr2:25659244	*DTNB*	intronic	rs6746082	4×10^{-7}		1.29 [1.17-1.42]	Illumina [422,839]	372
			3p22.1	chr3:41925398	*ULK4*	coding	rs1052501	2×10^{-8}		1.32 [1.20-1.45]		
			7p15.3	chr7:21938240	*DNAH11*	upstream	rs4487645	3×10^{-14}		1.38 [1.28-1.50]		

HM Subtype	Initial Sample Size	Replication Sample Size	Region	hg19 Chromosome and position	RefSeq gene location or genes within +/- 10 kb when intergenic	RefSeq gene region	SNP rsIDs	p-Value	p-Value note	OR [95% CI] (NR = not reported)	Platform [SNPs passing QC]	Reference
MM	2,335 European ancestry cases, 7,306 European ancestry controls	2,357 European ancestry cases, 3,684 European ancestry controls	10q25.2	chr10:112036975	*MXI1*	intronic	rs11195062	3×10^{-6}		1.13 [NR]	Illumina [414,804]	373
			17p11.2	chr17:16849139	*TNFRSF13B*	intronic	rs4273077	8×10^{-9}		1.26 [1.16-1.36]		
			2q37.1	chr2:232176541	*ARMC9*	downstream	rs7580717	5×10^{-6}		1.13 [NR]		
			22q13.1	chr22:39542292	*CBX7*	intronic	rs877529	8×10^{-16}		1.23 [1.17-1.29]		
			3q26.2	chr3:169492101	*MYNN*	coding	rs10936599	9×10^{-14}		1.26 [1.18-1.33]		
			6p21.33	chr6:31107258	*PSORS1C1*	intronic	rs2285803	1×10^{-10}		1.19 [1.13-1.26]		
MM	up to 1,660 European ancestry cases, 7,306 European ancestry controls		1p32.3	chr1:54375570	*DIO1*	intronic	rs2235544	6×10^{-7}	Hyperdiploid (HD) vs. controls	1.30 [1.17-1.44]	Illumina [414,804] (Imputed)	374
			11p14.3	chr11:21896880	intergenic	intergenic	rs11026318	9×10^{-6}	Non-HD without IgH translocations vs. controls	2.33 [1.60-3.38]		
			11q13.3	chr11:69457293	*CCND1*	intronic	rs1352075	2×10^{-6}	HD vs. non-HD	1.42 [1.23-1.64]		
			14q21.1	chr14:43403459	intergenic	intergenic	rs11157317	9×10^{-6}	HD vs. controls	1.29 [1.15-1.44]		
			17q21.1	chr17:38179492	*MED24*	intronic	rs2302777	8×10^{-7}	HD vs. controls	1.31 [1.18-1.46]		
			22q13.1	chr22:39519196	intergenic	intergenic	rs139371	2×10^{-9}	HD vs. controls	1.37 [1.24-1.52]		
			3q26.2	chr3:168916240	*MECOM*	intronic	rs9864370	9×10^{-6}	HD vs. non-HD	2.04 [1.49-2.80]		
			3p22.1	chr3:41786009	*ULK4*	intronic	rs6599175	1×10^{-9}	HD vs. controls	1.48 [1.30-1.68]		
			3p22.1	chr3:41996136	*ULK4*	intronic	rs2272007	2×10^{-9}	HD vs. controls	1.47 [1.30-1.67]		
			5q35.2	chr5:173777520	*LINC01411*	intronic	rs12659144	3×10^{-6}	HD vs. non-HD	1.67 [1.35-2.07]		
			7p15.3	chr7:21938240	*DNAH11*	upstream	rs4487645	1×10^{-9}	HD vs. controls	1.43 [1.27-1.60]		
			8p23.1	chr8:9395532	intergenic	intergenic	rs6601327	8×10^{-6}	HD vs. controls	1.27 [1.14-1.41]		
			8p23.1	chr8:9601699	*TNKS*	3'UTR	rs12545912	7×10^{-6}	HD vs. controls	1.30 [1.16-1.45]		
			8p23.1	chr8:9792662	intergenic	intergenic	rs2055729	8×10^{-7}	HD vs. controls	1.30 [1.17-1.45]		
			1q21.3	chr1:153640827	*ILF2*	intronic	rs7536700	4×10^{-6}	Any IgH translocation vs. controls	1.53 [1.28-1.84]		
			1q21.3	chr1:153640827	*ILF2*	intronic	rs7536700	9×10^{-6}	non t11;14 and t4;14 IgH translocations vs. controls	1.86 [1.42-2.46]		
			10p15.1	chr10:4676196	intergenic	intergenic	rs7086888	8×10^{-6}	non t11;14 and t4;14 IgH translocations vs. controls	1.87 [1.42-2.46]		
			10q22.1	chr10:72767409	intergenic	intergenic	rs10509328	1×10^{-6}	non t11;14 and t4;14 IgH translocations vs. controls	1.86 [1.45-2.39]		
			10q24.1	chr10:98639846	*LCOR*	intronic	rs17112190	9×10^{-7}	t(11;14) vs. controls	1.80 [1.42-2.27]		
			11q13.3	chr11:69462910	*CCND1*	exonic	rs9344	8×10^{-11}	t(11;14) vs. controls	1.82 [1.52-2.19]		
			11q13.3	chr11:69462910	*CCND1*	exonic	rs9344	2×10^{-11}	t(11;14) vs non t(11;14)	1.95 [1.61-2.38]		
			13q14.2	chr13:49466931	intergenic	intergenic	rs1449572	8×10^{-6}	t(4;14) vs. controls	1.73 [1.36-2.21]		
			14q22.3	chr14:57607067	intergenic	intergenic	rs7144018	9×10^{-6}	t(4;14) vs. controls	1.85 [1.41-2.44]		
			15q13.1	chr15:28488888	*HERC2*	intronic	rs8028689	7×10^{-6}	Any IgH translocation vs. controls	1.65 [1.33-2.05]		
			16q23.3	chr16:82787053	*CDH13*	3'UTR	rs8056064	2×10^{-6}	t(11;14) vs. controls	1.68 [1.36-2.08]		
			16q24.3	chr16:89448663	*ANKRD11*	intronic	rs3096299	4×10^{-6}	t(11;14) vs non t(11;14)	1.54 [1.28-1.86]		
			16q24.3	chr16:89471246	*ANKRD11*	intronic	rs2086824	2×10^{-6}	t(11;14) vs non t(11;14)	1.58 [1.31-1.91]		
			18q22.1	chr18:64952213	intergenic	intergenic	rs8099213	5×10^{-6}	non t(11;14) and t(4;14) IgH translocations vs. controls	1.52 [1.27-1.81]		
			2q31.1	chr2:174504924	intergenic	intergenic	rs13028485	2×10^{-6}	t(11;14) vs non t(11;14)	2.17 [1.57-2.99]		
			2p23.3	chr2:25613146	*DTNB*	intronic	rs7577599	6×10^{-6}	Any IgH translocation	1.43 [1.22-1.67]		

HM Subtype	Initial Sample Size	Replication Sample Size	Region	hg19 Chromosome and position	RefSeq gene location or genes within +/- 10 kb when intergenic	RefSeq gene region	SNP rsIDs	p-Value	p-Value note	OR [95% CI] (NR = not reported)	Platform [SNPs passing QC]	Reference
									vs. controls			
			2p23.3	chr2:25633242	*DTNB*	intronic	rs10180663	2×10^{-6}	t(11;14) vs. controls	1.64 [1.34-2.02]		
			22q13.1	chr22:39519196	intergenic	intergenic	rs139371	9×10^{-6}	non t(11;14) and t(4;14) IgH translocations vs. controls	1.50 [1.26-1.80]		
			3q13.33	chr3:121154193	*POLQ*	downstream	rs6800901	2×10^{-6}	any IgH translocation vs. controls	1.31 [1.17-1.47]		
			3q26.2	chr3:169477506	*TERC*	upstream	rs12638862	2×10^{-6}	any IgH translocation vs. controls	1.37 [1.20-1.56]		
			4q34.1	chr4:174655914	intergenic	intergenic	rs4521323	9×10^{-6}	any IgH translocation vs. controls	1.28 [1.15-1.43]		
			4q34.3	chr4:180968431	intergenic	intergenic	rs1994816	5×10^{-7}	non t(11;14) and t(4;14) IgH translocations vs. controls	1.76 [1.41-2.20]		
			5p14.3	chr5:22812264	*CDH12*	intronic	rs780179	9×10^{-7}	t(11;14) vs non t(11;14)	1.59 [1.32-1.92]		
			6q23.2	chr6:131563577	*AKAP7*	intronic	rs4629710	6×10^{-6}	non t(11;14) and t(4;14) IgH translocations vs. controls	1.60 [1.31-1.96]		
			8q12.1	chr8:60545664	intergenic	intergenic	rs4737547	3×10^{-6}	any IgH translocation vs. controls	1.30 [1.17-1.45]		
MPN	324 European ancestry cases, 2,999 European ancestry controls	NA	9p24.1	chr9:5070831	*JAK2*	intronic	rs10974944	4×10^{-20}		3.10 [2.40-4.00]	Affymetrix [62,775]	[375]
HL	393 European ancestry cases, 3,315 European ancestry controls	113 European ancestry cases, 214 European ancestry controls	6p21.32	chr6:32109979, chr6:32383108, chr6:32384721, chr6:32428285, chr6:32572251	*FKBPL, PRRT1, HCG23, HLA-DRA, HLA-DRB1*	intergenic	rs204999, rs9268528, rs9268542, rs6903608, rs2858870	8×10^{-18}		2.50 [NR]	Illumina [705,591]	[376]
			6p21.32	chr6:32109979, chr6:32383108, chr6:32384721, chr6:32428285, chr6:32572251	*FKBPL, PRRT1, HCG23, HLA-DRA, HLA-DRB1*	intergenic	rs204999, rs9268528, rs9268542, rs6903608, rs2858870	2×10^{-7}		1.70 [NR]		

Appendix 3.1 Phenol chloroform extraction of genomic DNA

Per volume (V) of diluted DNA ½V of phenol and ½V of chloroform were added and mixed using a vortex. Samples were then centrifuged in a bench top microcentrifuge at 13,000 rpm for 10 mins. The upper layer was removed to a new tube and the bottom layer discarded. 10 µL of 3 M sodium acetate and 1 µL glycogen was added and the volume was brought to 200 µL total with nuclease free H_2O. 200 µL of ice cold 100% molecular grade ethanol was added and the samples were mixed by vortex and incubated -20ºC 3 - 6 hours. Following incubation the samples were centrifuge at 15,000 rpm for 30 mins at 4ºC to pellet the DNA precipitate. The pellet supernatant was aspirated and discarded. A 200 µL wash of ice cold 80% molecular grade ethanol was added to each sample which were then mixed by vortex and centrifuged at 15,000 rpm for 5 mins at 4 ºC. The ethanol wash was aspirated from the DNA pellet. This was dried for 10 mins at room temperature to remove residual ethanol. The pellet was resuspend in a required volume of Tris-EDTA buffer (10 mM tris(hydroxymethyl)aminomethane) at pH 8.0 and 1mM ethylenediaminetetraacetic acid). Pellets in Tris-EDTA were dissolved using combinations of mixing by vortex, centrifugation and heat treatment at 52ºC for 5 mins.

Appendix 3.2 Custom awk script for consistency checking of mapped metadata and compliance to the SAM/BAM format during genome and exome alignment

```
#Set MAPQ to 0 and CIGAR to * for unmapped reads using
awk

...

bwa aln/mem ... | awk -v IFS="\t" -v OFS="\t" '{ if($$1
!~ /^@/ &&
and($$2,0x0004)) { $$5=0; $$6="*"; } ; print $$0}'
| samtools view

...
```

Appendix 3.3 Example SAMtools mpileup and GATK commands for one region.

SAMtools_variant_calling_and_quality_flagging.sh

```
# SAMtools mpileup multisample variant calling command used
samtools mpileup -C50 -D -g -S -uf hg19-sorted.fasta -r chr1:10000-
177417  LK0051_001.bam  LK0051_007.bam  LK0051_128.bam  LK0051_159.bam
LK0051_165.bam    LK0124_117.bam    LK0124_179.bam    LK0124_202.bam
LK0139_001.bam    LK0139_004.bam    LK0139_005.bam    LK0153_003.bam
LK0153_004.bam    LK0153_029.bam    LK0153_078.bam    LK0153_079.bam
LK0153_080.bam    LK0153_084.bam    LK0153_086.bam    LK2042_003.bam
LK2042_005.bam    LK2042_006.bam    LK2042_018.bam    LK2042_231.bam
LK2042_232.bam    LK2042_257.bam    LK2042_258.bam    LK2042_259.bam
LK2042_281.bam LK2042_290.bam LK2042_300.bam | bcftools view -vcg -
> Region_1.vcf

# Adding variant quality flags to each variant in each individual
using GATK
/usr/java/latest/bin/java                -Xmx2g                -jar
/usr/local/software/gatk/default/GenomeAnalysisTK.jar \
   -R hg19-sorted.fasta \
   -T VariantFiltration \
   -o Region_1.filter.vcf \
   --variant Region_1.vcf \
     --genotypeFilterExpression "DP >= 10 && GQ >= 20" \
     --genotypeFilterName "High_Sample_Confidence" \
     --genotypeFilterExpression "DP < 10 && GQ < 20" \
     --genotypeFilterName "Low_Sample_Confidence"

# Extracting only variants from the VCF file that have been tagged
with High_Sample_Confidence in at least one individual
egrep     "^#|High_Sample_Confidence"     Region_1.filter.vcf     >
Region_1.High_Sample_Confidence.vcf


# Conversion of VCF file to ANNOVAR input using a custom script
convert_VCF_to_ANNOVAR_input.sh   Region_1.High_Sample_Confidence.vcf
> Region_1.High_Sample_Confidence.annovar
```

Appendix 3.4 Custom VCF to ANNOVAR input conversion script

convert_VCF_to_ANNOVAR_input.sh

```
# Custom script for converting VCF to an input format compatible with ANNOVAR

#!/usr/bin/awk -f

# Purpose of this script is to convert multisample VCF files into input format ANNOVAR

#awk -v FS="\t"
BEGIN {
     FS="\t";
     OFS="\t";
     # Print header
     print
"#CHROM"OFS"START"OFS"END"OFS"REF"OFS"ALT"OFS"INFO"OFS"LK0051-
001"OFS"LK0051-007"OFS"LK0051-128"OFS"LK0051-159"OFS"LK0051-
165"OFS"LK0124-117"OFS"LK0124-179"OFS"LK0124-202"OFS"LK0139-
001"OFS"LK0139-004"OFS"LK0139-005"OFS"LK0153-003"OFS"LK0153-
004"OFS"LK0153-029"OFS"LK0153-078"OFS"LK0153-079"OFS"LK0153-
080"OFS"LK0153-084"OFS"LK0153-086"OFS"LK2042-003"OFS"LK2042-
005"OFS"LK2042-006"OFS"LK2042-018"OFS"LK2042-231"OFS"LK2042-
232"OFS"LK2042-257"OFS"LK2042-258"OFS"LK2042-259"OFS"LK2042-
281"OFS"LK2042-290"OFS"LK2042-
300"OFS"COUNT_ALT"OFS"HAS_ALT"OFS"HAS_REF"
}

$0 ~/^#CHROM/ {
     for(i=10;i<NF+1;i++){
          sample_name[i]=$i;
     }
}

$0 !~ /^#/ {
     refl=length($4);
     printf
"%s"OFS"%i"OFS"%i"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"O
FS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s
"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"
%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s"OFS"%s",$1,$2,$2+refl-
1,$4,$5,$8,$10,$11,$12,$13,$14,$15,$16,$17,$18,$19,$20,$21,$22,$23,$
24,$25,$26,$27,$28,$29,$30,$31,$32,$33,$34,$35,$36,$37,$38,$39,$40

     # Parse the format column
     n=split($9,format,":")
     for (i=1;i<n;i++){
          if (format[i] == "GT") {
               gt_field=i;
               break;
          }
```

```
    }
    sum=0;
  with="";
  without="";
    for (i=10;i<NF+1;i++) {
        split($i,genotypes,":");
        gt=genotypes[gt_field];
        if ( gt == "0/0" ) {
            c=0;
            without=without","sample_name[i]
        } else if ( gt == "0/1" ) {
            c=1;
            with=with","sample_name[i];
        } else if ( gt == "1/1" ) {
            c=1;
            with=with","sample_name[i];
        };
        sum=sum+c;
    }
    sub(",","",with);
    sub(",","",without);
    printf OFS"%s"OFS"%s"OFS"%s\n",sum,with,without
}
```

Appendix 3.5 pVAAST code

```
# This parameter file is optimized for rare Mendelian diseases with
# large pedigrees. If the penetrance and prevalence is known, please
# also modify "penetrance_lower_bound", "penetrance_upper_bound" and
# "max_prevalence_filter" parameters.

#------------------------Basic options-----------------------------
input_ped_cdr_files:  LK0051_pedigree.ped  LK0051_Target_variants.cdr
# space separated list of ped and cdr files

pedigree_representatives:   LK0051-128

unknown_representatives:    no
additional_cases:
inheritance_model:          dominant
penetrance_lower_bound:     0
penetrance_upper_bound:     1


#--------------------Performance Tuning -------------------------

informative_site_selection:     3    # 1 stands for selecting sites
based on CLRT score;
                        # 2 for LOD score; 3 for LOD + CLRT score
simulate_genotyping_error:  yes #  Use  this  option  will  prevent
inheritance error to
                        #  achieve  0  p-values  in  fully  sequenced
trios. For de novo
                        #   mutations,   set   "inheritance_model"   to
"dominant" and
                        # this option to "yes"
genotyping_error_rate:      1e-4 #  Set  the  genotyping  error  rate.
Essential for de novo
                        # mutations.



#----------------Gene and Variant Filtering ------------------------

max_prevalence_filter:      0.1
lod_score_filter:           no
clrt_score_filter:          no
nocall_filter:              no
nocall_filter_cutoff:       no
inheritance_error_filter:   no



#----------------------Developer Options--------------------------
clrt_randomization_round:   100
locus_heterogeneity_penalty:    0
incomplete_penetrance_penalty:   0
mcmc_use_functional_score:    yes
```

Appendix 3.6 Representative genome (LK2042-003) and exome (LK2042-005) FastQC reports and LK2042-005 Qualimap report

FastQC Report

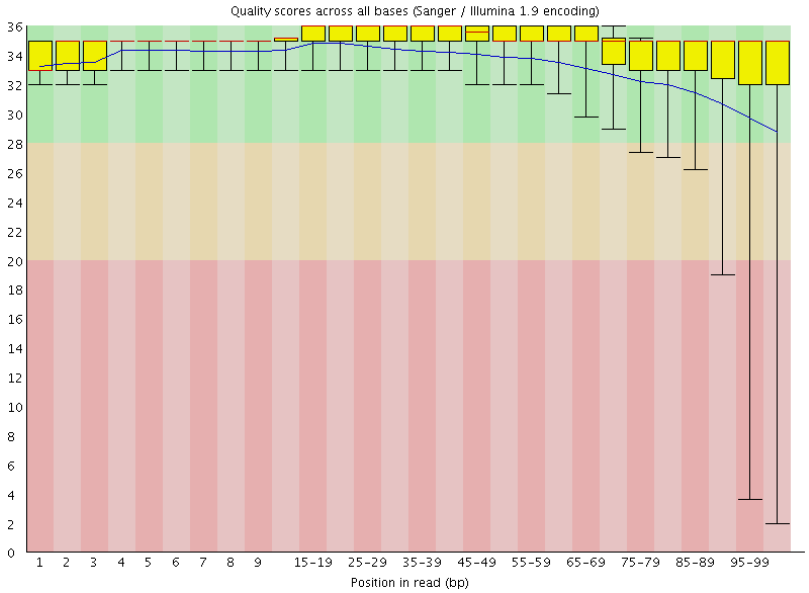
Tue 15 Jul 2014
LK2042_003.bam

Summary

Basic Statistics

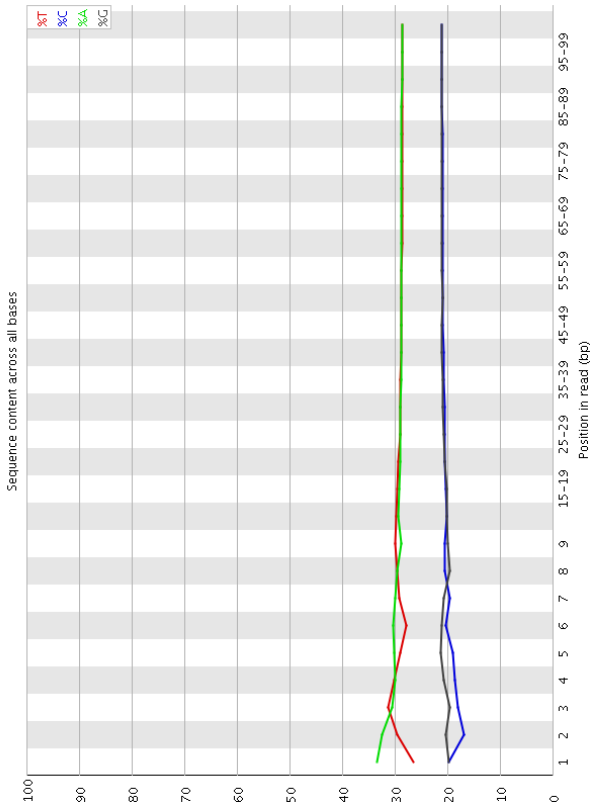
Measure	Value
Filename	LK2042_003.bam
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1371921146
Filtered Sequences	0
Sequence length	100
%GC	41

Per base sequence quality

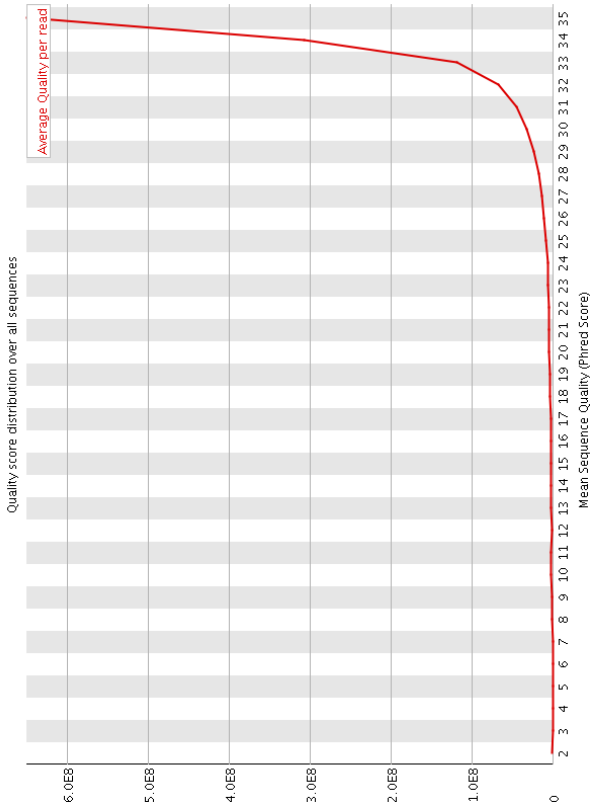


Per sequence quality scores

LK2042-003 FastQC Report

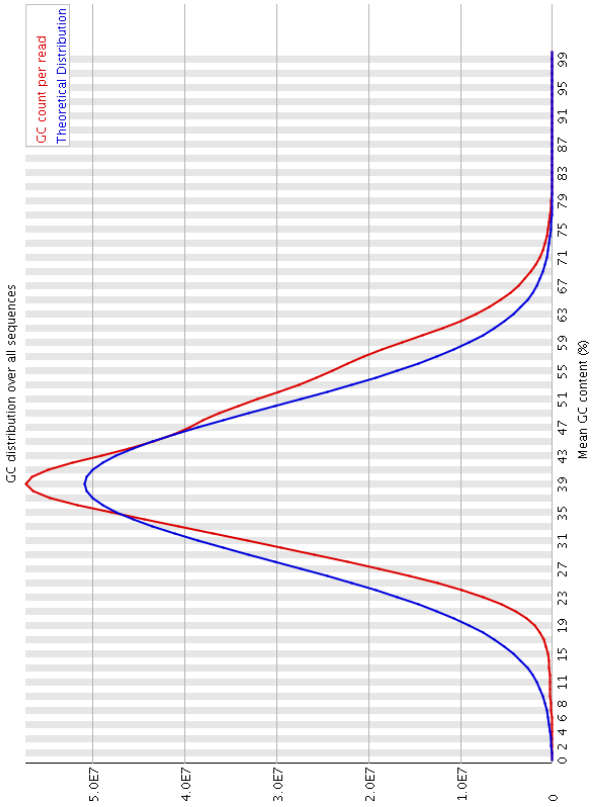


Per base GC content

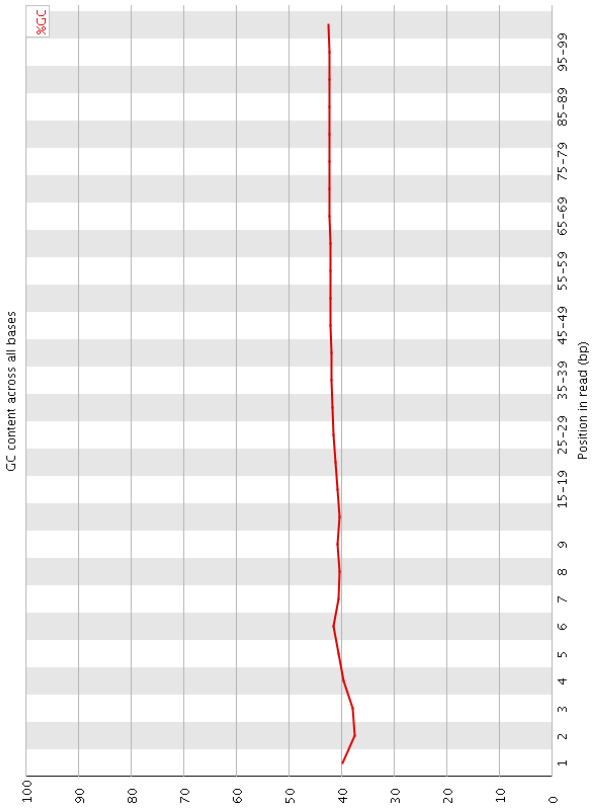


Per base sequence content

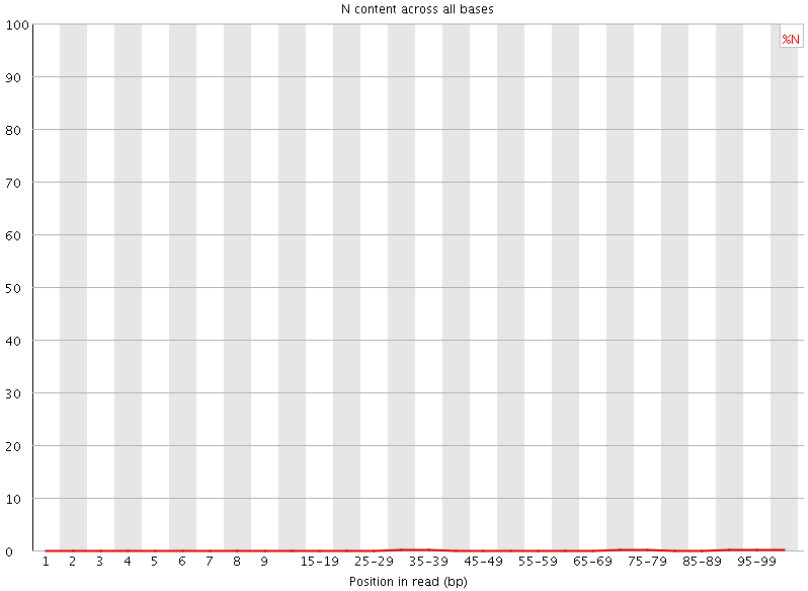
LK2042-003 FastQC Report

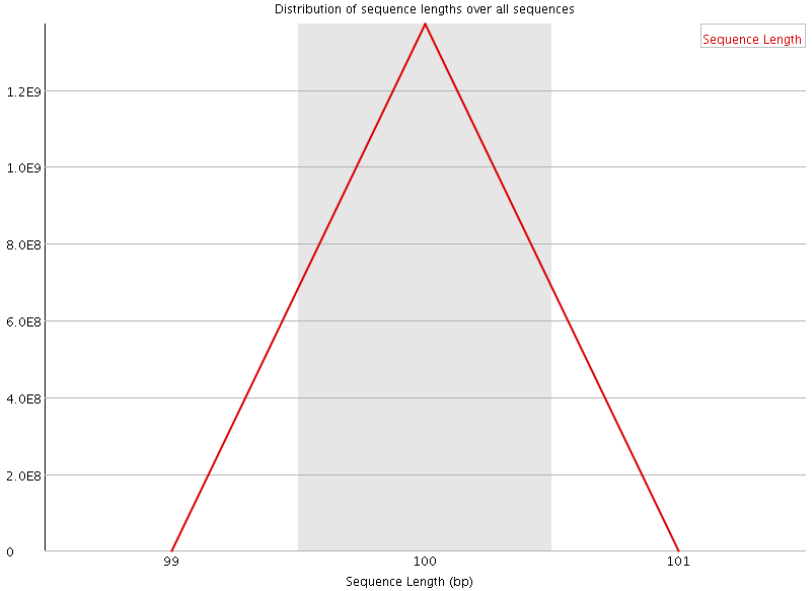


Per base N content



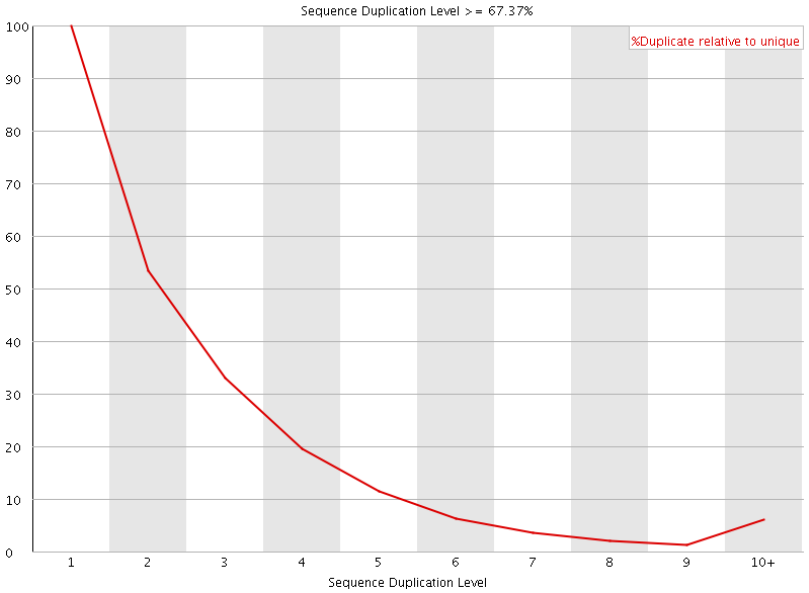
Per sequence GC content

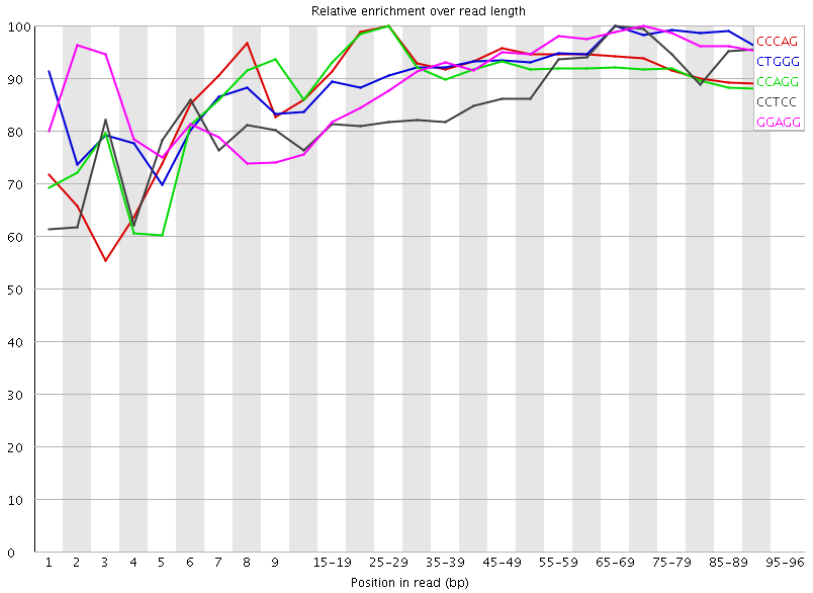




Sequence Length Distribution

Sequence Duplication Levels





Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
CCCAG	248202680	3.4221373	3.7414453	25-29
CTGGG	240660065	3.276855	3.5289	65-69
CCAGG	224033065	3.0633762	3.3857577	25-29
CCTCC	216034515	3.0156748	3.4531605	65-69
GGAGG	224079890	3.0136	3.2821248	70-74

Overrepresented sequences

No overrepresented sequences

Kmer Content

Produced by FastQC (version 0.10.0)

FastQC Report

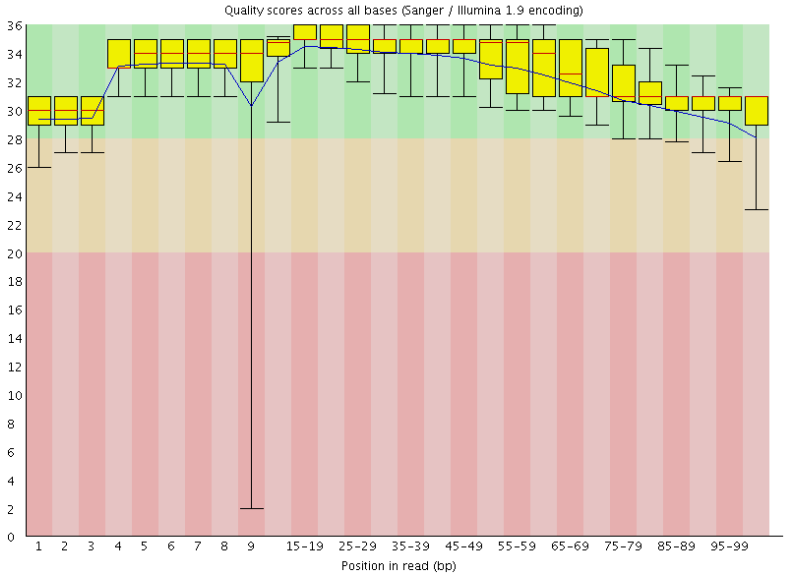
Fri 7 Mar 2014
LK2042-005.recalibrated.bam

Summary

Basic Statistics

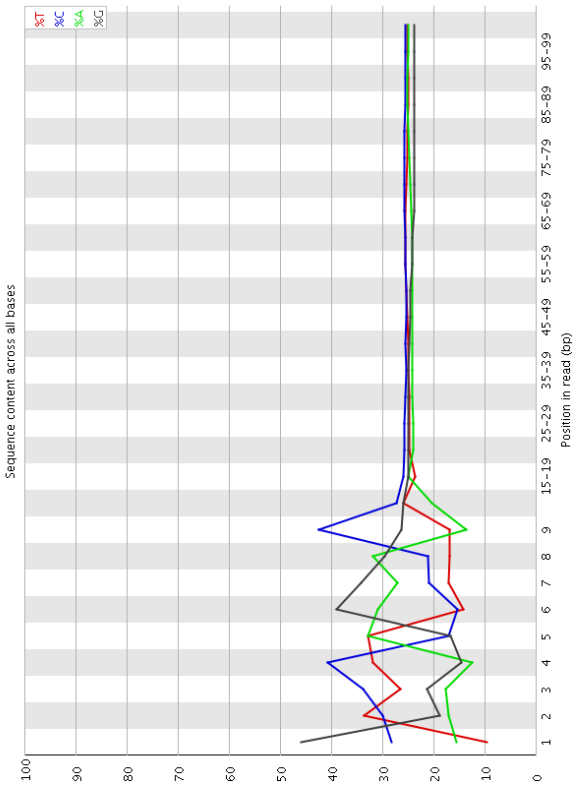
Measure	Value
Filename	LK2042-005.recalibrated.bam
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	52587692
Filtered Sequences	0
Sequence length	100
%GC	50

Per base sequence quality

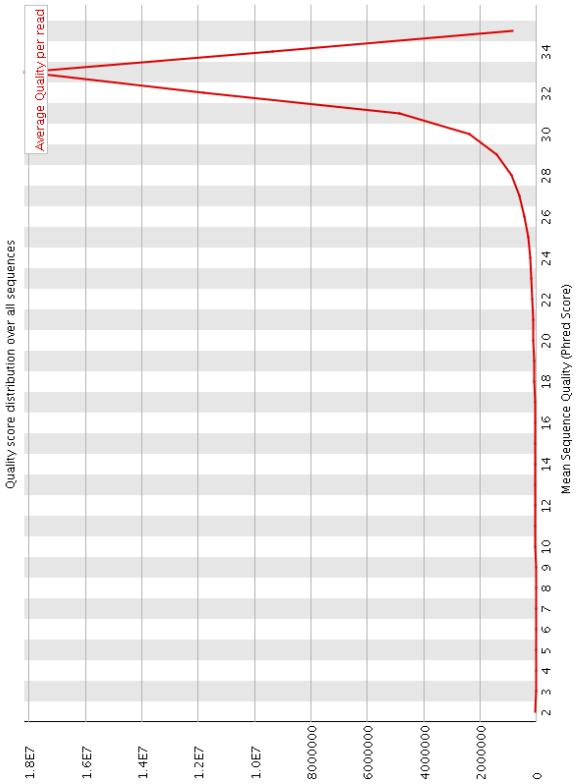


Per sequence quality scores

LK2042-005 FastQC Report



Per base GC content



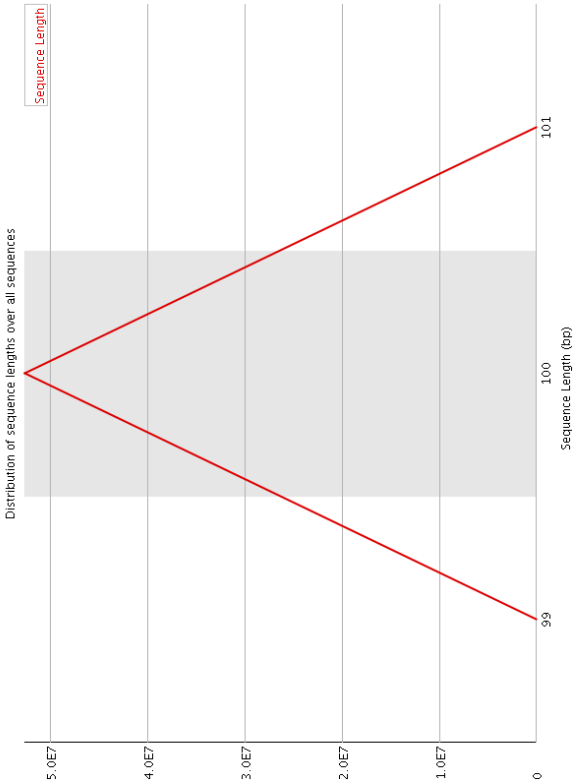
Per base sequence content

LK2042-005 FastQC Report

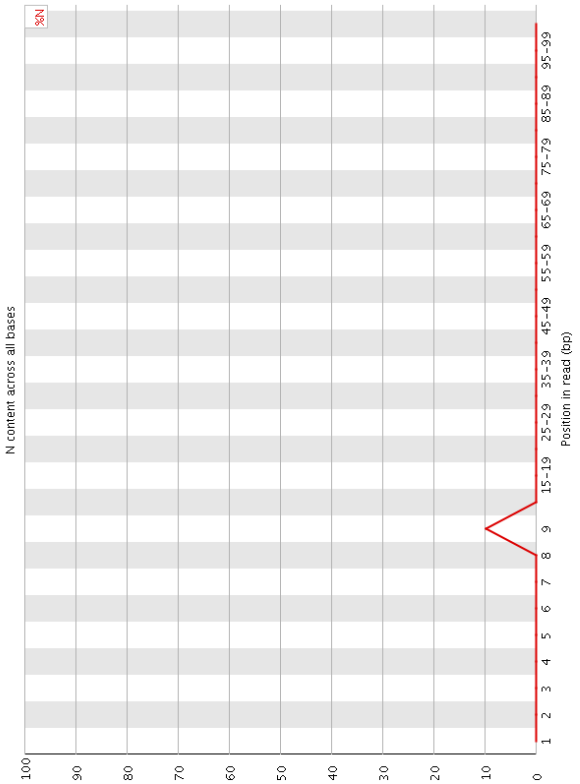
Per base N content

Per sequence GC content

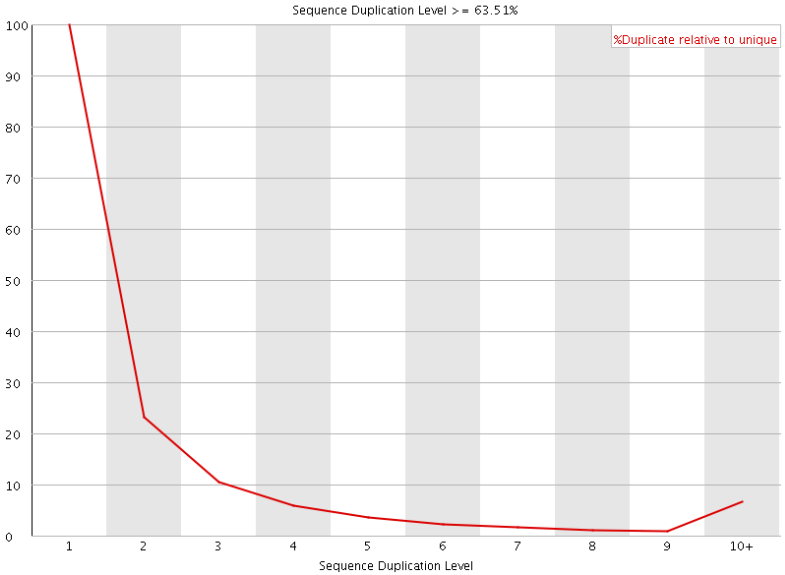
LK2042-005 FastQC Report

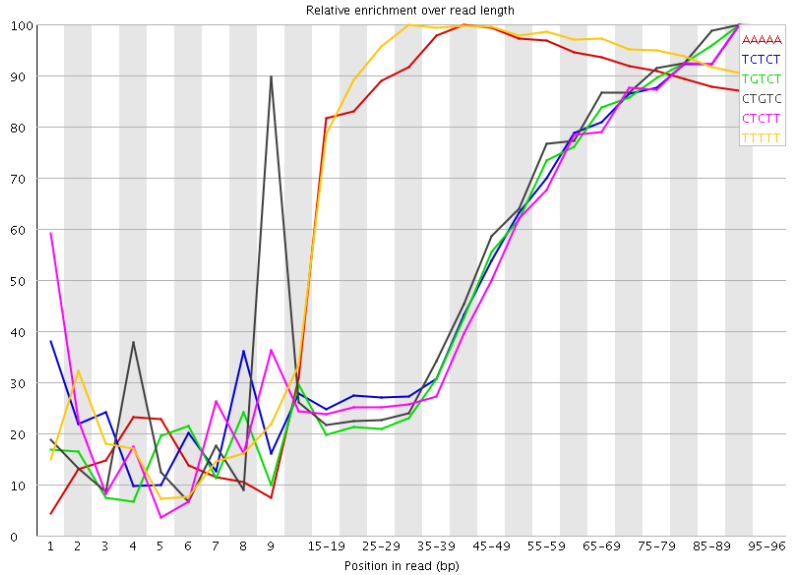


Sequence Duplication Levels



Sequence Length Distribution





Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
AAAAA	16365570	3.8904545	4.752976	40-44
TCTCT	19661020	3.7280314	6.524641	90-94
TGTCT	18174005	3.6101713	6.4494214	90-94
CTGTC	18657380	3.559149	6.1169095	90-94
CTCTT	18566405	3.5204759	6.2970695	90-94
TTTTT	15708925	3.2298563	3.8096044	30-34
GTCTC	16887720	3.2215624	6.024262	90-94
ACACA	15267700	3.1583965	6.12459	95-96
TTATA	13850645	3.0180032	6.7553287	95-96
CATCT	15150535	2.9573863	5.880235	95-96
TCTTA	14447300	2.9366238	6.214357	95-96
ATACA	13211740	2.8459983	6.3200173	95-96
CACAT	13794435	2.7719824	5.762486	90-94
CTTAT	13421395	2.7280936	6.1234126	95-96
CCAGG	13796155	2.7257252	12.960482	3
CCTGG	14041365	2.694802	12.055865	3

Overrepresented sequences

No overrepresented sequences

Kmer Content

LK2042-005 FastQC Report

AGAGG	11767010	2.6108496	5.8523254	5
ACATC	12979145	2.6081502	5.699054	90-94
GCTGC	13575905	2.6054718	5.966066	8
CTGGG	12929745	2.5996275	11.577891	4
GGAGG	11873555	2.5746133	6.140871	6
CCTGT	13349205	2.5465424	5.382407	8
GGCTG	12498500	2.5129225	9.318972	7
TATAC	11897815	2.4896333	6.0794497	95-96
CTGCC	13573530	2.4866045	5.3937154	9
CCCAG	12991680	2.4501114	10.793731	2
AGGAG	10938170	2.4269478	7.785143	5
AGCTG	11762760	2.4199972	6.904563	7
TACAC	11925165	2.396354	5.519588	95-96
GCTGG	11841200	2.380767	7.2047634	3
TGCTG	11847750	2.3677454	7.661318	2
GCAGG	11121110	2.3018436	8.275038	3
CCCTG	12557180	2.3004143	9.01237	2
CAGGA	10853255	2.2986465	10.9767	4
TCCTG	11927360	2.2753065	9.529562	2
CCAGC	12058605	2.2741418	6.624734	3
GCCTG	11837915	2.2719188	7.164958	7
CAGCA	11000810	2.223992	7.4641266	4
AGCAG	10420435	2.206978	6.2261553	7
CTGGA	10591515	2.1790326	9.5865135	4
CAGGG	10480995	2.1693528	10.800362	4
GGCAG	10435825	2.1600037	6.499545	7
GCTGT	10747990	2.1479611	5.572814	8
CTCCT	11752740	2.140085	6.503704	1
TGGAG	9927290	2.1396363	7.5579515	5
TCCAG	10891095	2.1388168	9.796166	2
TGCAG	10380055	2.135528	8.291577	2
AGAAG	9317955	2.1155381	5.1607084	5
GCCCA	11195925	2.1114485	7.2031155	1
CTGCT	10994220	2.097297	6.632923	9
CCTGC	11387315	2.0861006	6.851586	3
CTGCA	10528895	2.0676873	7.656156	4
GGGAG	9505545	2.0611439	5.616136	7
CAGAG	9698835	2.054148	8.513958	4
ATCTC	10423750	2.034717	5.7959285	95-96

GAGAG	9081425	2.0149755	5.6946363	7
CTCTG	10374420	1.9790621	7.7614	2
CTCCA	10547295	1.9771496	7.2802763	1
CAGGC	9992835	1.9742986	7.213253	4
TGGGG	9339265	1.96715	6.809883	5
GGAGA	8814660	1.9557858	5.835868	6
ACCTG	9955110	1.9550062	5.1810155	7
AGGTG	8946230	1.9281877	5.4912615	7
GCCAG	9379590	1.8531387	5.3690476	2
CTGTG	9228730	1.8443403	9.033192	9
CACAG	9112200	1.842179	6.925774	7
CAGAA	8483785	1.8385961	5.5899625	4
CTGGC	9549030	1.8326384	5.6682243	4
GGCCT	9539900	1.8308864	6.3165402	6
CTGAG	8869520	1.82476	6.498092	4
TTCTG	8921570	1.7722231	5.2305837	2
CTCAG	8949485	1.7575192	6.395504	2
AGACA	8063045	1.7474136	5.275386	6
AGGCC	8769345	1.7325717	6.2489433	5
CCCCA	9606910	1.7294204	6.137699	1
GGCCA	8701965	1.7192594	6.021191	1
GGGCC	8870455	1.7127163	5.905417	6
TGGCC	8722770	1.6740638	6.017762	5
CCAGA	8212695	1.6603292	6.7276287	3
CCTCT	9110465	1.6589468	5.32803	1
TGGGC	8223465	1.6533927	7.501728	5
TCTGG	8237390	1.6462234	6.829306	3
GGCCC	8887785	1.6380589	5.146483	6
GGGCT	8098160	1.6281991	7.057235	6
GGGCA	7569875	1.5668101	5.1152353	6
GAGCT	7556870	1.5547035	5.1855545	6
AGGGG	7168320	1.5543493	5.2191586	5
CACTG	7905745	1.5525475	7.2845306	7
CAGTG	7495090	1.5419934	6.045094	9
AGGCT	7488905	1.5407208	5.3954067	6
AGGGC	7381675	1.5278565	6.8885336	5
TCAGG	7377570	1.5178155	6.1240997	3
ACAGA	6905585	1.4965702	5.8322077	8
GCCCT	8124150	1.4883049	6.1565366	1

LK2042-005 FastQC Report

ACAGG	7026060	1.4880722	5.17989	8
CCTGA	7564650	1.4855623	5.701599	3
CTTGG	7095730	1.4180653	5.130783	3
AGAGC	6616800	1.4013937	5.532751	5
ACAGC	6884025	1.3917172	5.5517216	8
GGGTG	6558935	1.3815229	5.15317	7
GACAG	6278435	1.3297303	6.990361	7
GTGTG	6297340	1.3184385	7.204546	1
GCTCT	6776940	1.2927936	5.091988	1
GGACA	5987200	1.2680486	6.5127673	6
AGGAC	5785790	1.2253913	7.3344855	5
GTGCT	5984960	1.1960806	6.471434	1
CACAC	6184040	1.193376	5.782183	5
GTCCT	6167745	1.1765813	6.832936	1
ACTGT	5593585	1.1438627	6.378711	8
AGCAC	5515130	1.114973	6.0113354	5
TGGAC	5264765	1.0831401	6.6769266	5
GTCCA	5407890	1.0620131	6.757715	1
ACTGC	5404710	1.0613887	5.069821	8
GACTG	5070980	1.043272	6.006467	7
GTGCA	4917790	1.0117557	6.6383185	1
GGACT	4760670	0.9794307	5.839325	6
TGCAC	4938750	0.9698824	5.954702	5
ACAGT	4502000	0.94775397	5.5113263	8
GCACT	4555105	0.8945414	5.2468796	6
ACACT	4264180	0.8568841	5.1005435	6
GTGTA	3731540	0.7994206	5.9335675	1

Produced by FastQC (version 0.10.0)

Qualimap Analysis Results

BAM QC analysis
Generated by Qualimap v.2.0
2014/10/26 05:20:48

1. Input data & parameters (inside of regions)

1.1. QualiMap command line

```
qualimap bamqc -bam /master/nickb/exomes/LK2042_005.bam -gff truseq_exome_targeted_regions.hg19.bed.chr -nw 400 -hm 3
```

1.2. Alignment

BAM file:	/master/nickb/exomes/LK2042_005.bam
Program:	GATK IndelRealigner (1.4-37-g0b29d54)
Size of a homopolymer:	3
Number of windows:	400
Analysis date:	Sun Oct 26 05:20:46 CDT 2014
Draw chromosome limits:	no

1.3. GFF region

GFF file:	truseq_exome_targeted_regions.hg19.bed.chr
Library protocol:	non-strand-specific
Outside statistics:	no

2. Summary (inside of regions)

2.1. Globals

Reference size	3,137,161,264
Number of reads	52,587,692
Mapped reads	40,827,929 / 77.64%
Unmapped reads	11,759,763 / 22.36%
Paired reads	40,827,929 / 77.64%
Mapped reads, only first in pair	20,384,075 / 38.76%
Mapped reads, only second in pair	20,443,854 / 38.88%
Mapped reads, both in pair	40,418,432 / 76.86%
Mapped reads, singletons	409,497 / 0.78%
Read min/max/mean length	100 / 100 / 100

2.2. Globals (inside of regions)

Regions size/percentage of reference	62,085,286 / 1.98%
Mapped reads	25,696,259 / 48.86%
Mapped reads, only first in pair	12,800,517 / 24.34%
Mapped reads, only second in pair	12,895,742 / 24.52%
Mapped reads, both in pair	25,446,212 / 48.39%
Mapped reads, singletons	250,047 / 0.48%
Correct strand reads	0 / 0%
Clipped reads	152,579 / 0.29%
Duplication rate	39.78%

2.3. ACGT Content (inside of regions)

Number/percentage of A's	511,324,654 / 24.13%
Number/percentage of C's	548,216,449 / 25.87%
Number/percentage of T's	512,216,502 / 24.17%
Number/percentage of G's	547,434,773 / 25.83%
Number/percentage of N's	0 / 0%
GC Percentage	51.7%

2.4. Coverage (inside of regions)

Mean	34.17
Standard Deviation	39.55

2.5. Mapping Quality (inside of regions)

Mean Mapping Quality	44.1

2.6. Insert size (inside of regions)

Mean	7,203.64
Standard Deviation	707,739.18
P25/Median/P75	121 / 156 / 207

2.7. Mismatches and indels (inside of regions)

General error rate	0.65%
Mismatches	13,265,447
Insertions	377,965
Deletions	231,240

Homopolymer indels	33.35%

2.8. Chromosome stats (inside of regions)

Name	Length	Mapped bases	Mean coverage	Standard deviation
chr1	6235259	220476726	35.36	44.38
chr2	4375163	144101099	32.94	41.44
chr3	3699926	121099217	32.73	39.54
chr4	2560752	74939836	29.26	26
chr5	3002851	99789533	33.23	34.8
chr6	3261463	91397027	28.02	26.56
chr7	2966092	111226397	37.5	46.02
chr8	2221183	79576381	35.83	87.11
chr9	2477630	91782771	37.04	37.16
chr10	2543587	84852838	33.36	30.93
chr11	3490061	125361095	35.92	34.76
chr12	3320688	110488544	33.27	29.99
chr13	1243459	37784021	30.39	25.38
chr14	1947408	64987412	33.37	34.72
chr15	2115182	75521843	35.7	36.47
chr16	2433884	94729167	38.92	36.41
chr17	3440233	134036944	38.96	36.48
chr18	1072537	31558027	29.42	24.76
chr19	3568108	141440536	39.64	38.1
chr20	1581628	56939649	36	33.51

chr21	668868	25670200	38.38	46.91
chr22	1390435	57896116	41.64	44.42
chrX	2372247	43449051	18.32	17.52
chrY	96642	2302514	23.83	30.75
chrM	0	0	0	0
chr1_gl000191_random	0	0	0	0
chr1_gl000192_random	0	0	0	0
chr4_ctg9_hap1	0	0	0	0
chr4_gl000193_random	0	0	0	0
chr4_gl000194_random	0	0	0	0
chr6_apd_hap1	0	0	0	0
chr6_cox_hap2	0	0	0	0
chr6_dbb_hap3	0	0	0	0
chr6_mann_hap4	0	0	0	0
chr6_mcf_hap5	0	0	0	0
chr6_qbl_hap6	0	0	0	0
chr6_ssto_hap7	0	0	0	0

chr7_gl000195_random	0	0	0	0
chr8_gl000196_random	0	0	0	0
chr8_gl000197_random	0	0	0	0
chr9_gl000198_random	0	0	0	0
chr9_gl000199_random	0	0	0	0
chr9_gl000200_random	0	0	0	0
chr9_gl000201_random	0	0	0	0
chr11_gl000202_random	0	0	0	0
chr17_ctg5_hap1	0	0	0	0
chr17_gl000203_random	0	0	0	0
chr17_gl000204_random	0	0	0	0
chr17_gl000205_random	0	0	0	0
chr17_gl000206_random	0	0	0	0
chr18_gl000207_random	0	0	0	0
chr19_gl000208_random	0	0	0	0

chr19_gl000209_random	0	0	0	0
chr21_gl000210_random	0	0	0	0
chrUn_gl000211	0	0	0	0
chrUn_gl000212	0	0	0	0
chrUn_gl000213	0	0	0	0
chrUn_gl000214	0	0	0	0
chrUn_gl000215	0	0	0	0
chrUn_gl000216	0	0	0	0
chrUn_gl000217	0	0	0	0
chrUn_gl000218	0	0	0	0
chrUn_gl000219	0	0	0	0
chrUn_gl000220	0	0	0	0
chrUn_gl000221	0	0	0	0
chrUn_gl000222	0	0	0	0
chrUn_gl000223	0	0	0	0

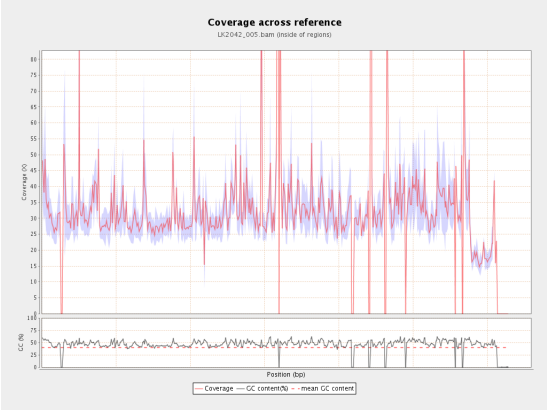
chrUn_gl000224	0	0	0	0
chrUn_gl000225	0	0	0	0
chrUn_gl000226	0	0	0	0
chrUn_gl000227	0	0	0	0
chrUn_gl000228	0	0	0	0
chrUn_gl000229	0	0	0	0
chrUn_gl000230	0	0	0	0
chrUn_gl000231	0	0	0	0
chrUn_gl000232	0	0	0	0
chrUn_gl000233	0	0	0	0
chrUn_gl000234	0	0	0	0
chrUn_gl000235	0	0	0	0
chrUn_gl000236	0	0	0	0
chrUn_gl000237	0	0	0	0
chrUn_gl000238	0	0	0	0

chrUn_gl000239	0	0	0	0
chrUn_gl000240	0	0	0	0
chrUn_gl000241	0	0	0	0
chrUn_gl000242	0	0	0	0
chrUn_gl000243	0	0	0	0
chrUn_gl000244	0	0	0	0
chrUn_gl000245	0	0	0	0
chrUn_gl000246	0	0	0	0
chrUn_gl000247	0	0	0	0
chrUn_gl000248	0	0	0	0
chrUn_gl000249	0	0	0	0

BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

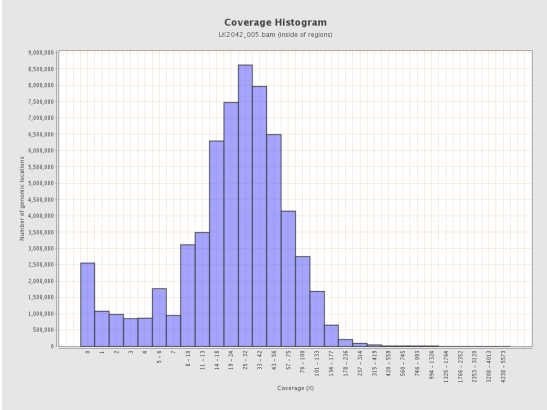
3. Results : Coverage across reference (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

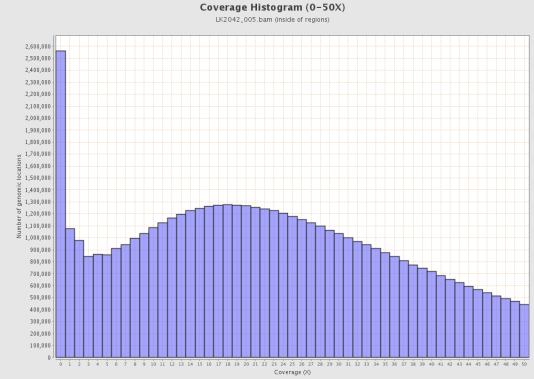
4. Results : Coverage Histogram (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACION

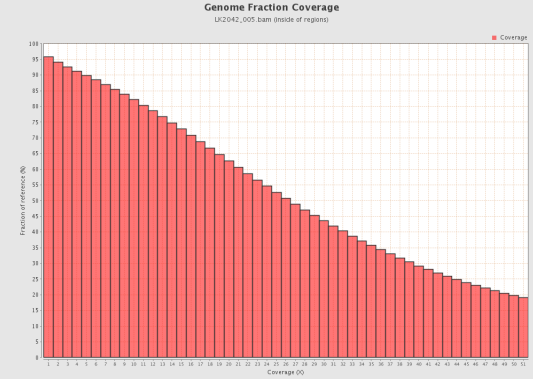
5. Results : Coverage Histogram (0-50X) (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACION

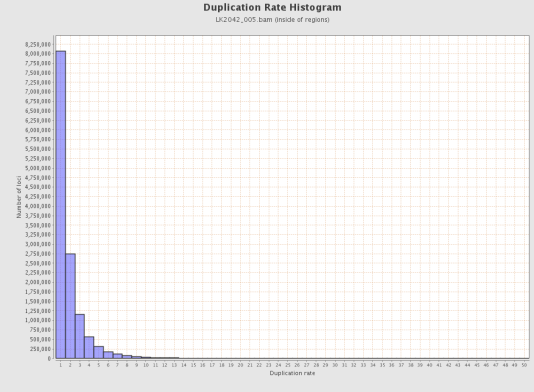
6. Results : Genome Fraction Coverage (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACION

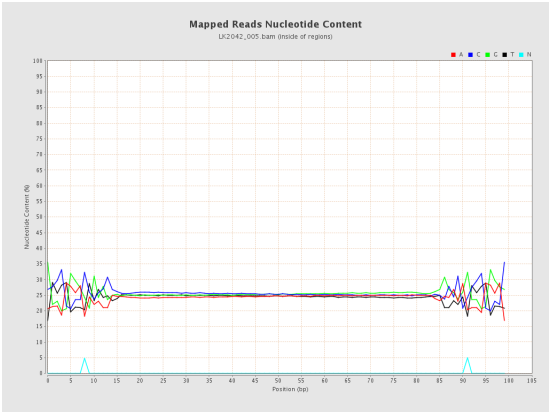
7. Results : Duplication Rate Histogram (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACION

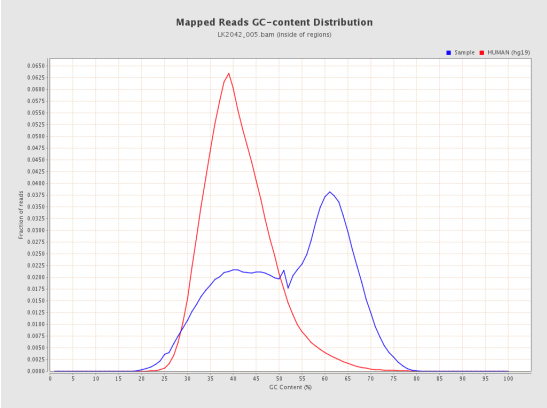
8. Results : Mapped Reads Nucleotide Content (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

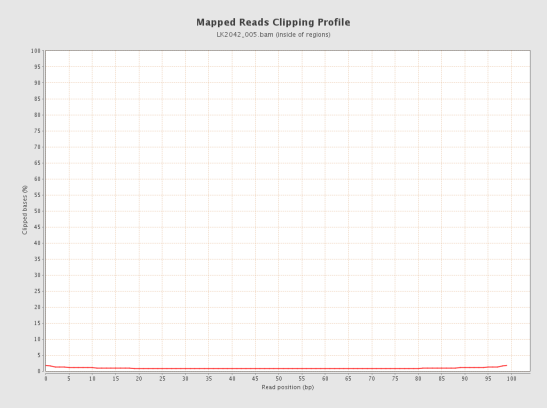
9. Results : Mapped Reads GC-content Distribution (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

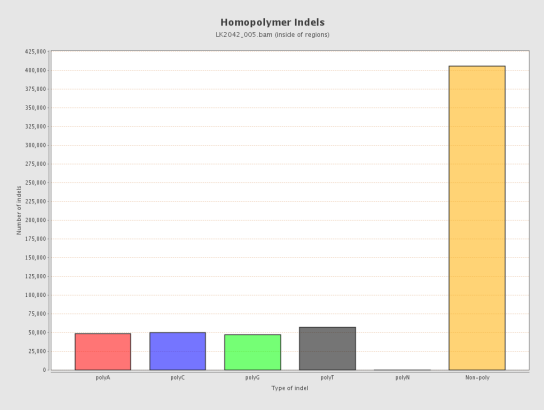
10. Results : Mapped Reads Clipping Profile (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACION

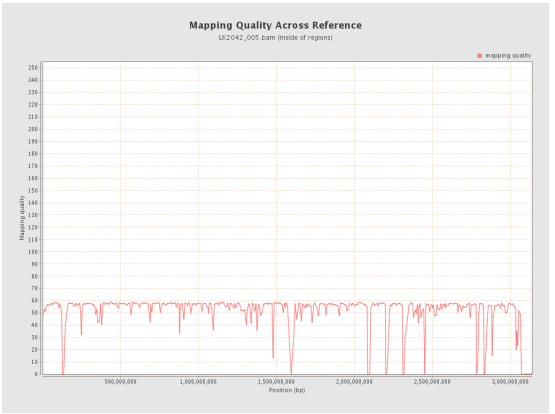
11. Results : Homopolymer Indels (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACION

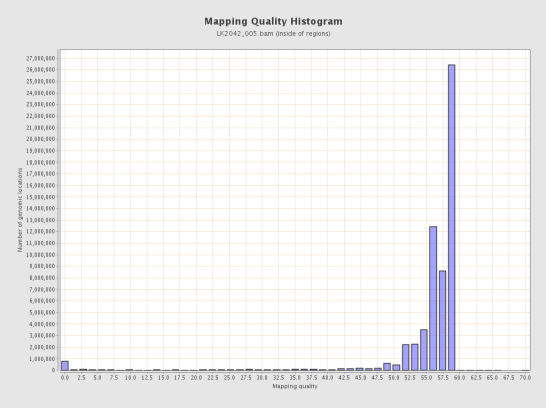
12. Results : Mapping Quality Across Reference (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

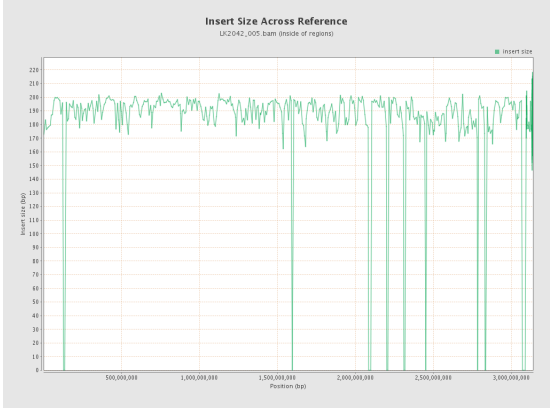
13. Results : Mapping Quality Histogram (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

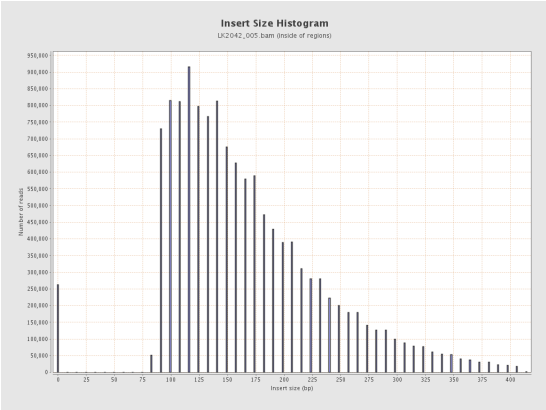
14. Results : Insert Size Across Reference (inside of regions)



BioinformaticsandGenomics

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

15. Results : Insert Size Histogram (inside of regions)



Appendix 4.1 Primer sequences and annealing temperatures

Gene	Variant	Forward Primer 5' to 3'	Reverse Primer 5' to 3'	Annealing Temp	PCR*
TNFSF9	chr19:6534728 G>C	ctgacatgttcggtgctcag	cctggcctcagtgtgaagat	62ºC	GTG
TDP2	chr6:24658126 C>T	tgctttttcctctctatgtccca	ggtatgttgtgggatttccttcat	62ºC	MyTaq
MMP8	chr11:102585288 T>G	gcttgaaggtttgtaagcaggg	aggcaaccaatactgggctc	62ºC	MyTaq
SPHK2	chr19:49132198 A>C	acccactggacctgctct	ttcggagacgggtgagtgta	62ºC	MyTaq
LRP5	chr11:68181292 C>A	cctggctgtgcctttgct	cccagaaccagcctgatcta	62ºC	GTG
NOTCH1	chr9:139417464 T>G	gtctggggaactcgccatc	tgtctccagggaaatcgtgc	62ºC	MyTaq
RIPK2	chr8:90796371 GTAA>G	acctatgtgacaagaagaaaatgga	acagggacagaactcaatgtca	62ºC	MyTaq
NAT10	chr11:34152973 C>A	ccttgtctcagggcatggtt	atgttgaggcaatccaggca	62ºC	MyTaq
PABPC1	chr8:101721839 C>A	tacctgtgggatagctgcca	tgacctggctagggtaatgc	62ºC	MyTaq
MET	chr7:116381047 A>G	ttgggctgcagacatttcca	acttcaggtgcaggacttgg	62ºC	MyTaq
DUSP10	chr1:221874839 TG>T	ttgcaggagaggatgacagc	tgccttcactggcttcttgt	62ºC	MyTaq
RARS	chr5:167929060 CCTT>C	tgtgtaccttaccattttctttccc	agagtgcctacctctatcttca	62ºC	MyTaq
HAL	chr12:96371767 A>G	aggaccagcttagttctgcat	tgttgactccctctccacca	62ºC	MyTaq
PEX6	chr6:42934551 G>A	cacctccaaggacttggtct	ggtccctgaaggttcctg	62ºC	GTG
GIT1	chr17:27904190 G>A	gctcgaggttgtctgaggac	cctgcgtcagtcataagggg	62ºC	MyTaq
NID2	chr14:52509033 G>A	aggcctccaattggtgtgag	ggaggaggggagaaaacaga	62ºC	MyTaq
STT3B	chr3:31659458 T>G	gtgtttctctggaagtgggttc	gaagtgttgcccaagaactga	62ºC	MyTaq
NF1	chr17:29508805 T>G	aggcattttggaactgggtaga	ggaagagggttgggcctaattt	62ºC	MyTaq

*GTG = GoTaq Green PCR master mix, MyTaq = MyTaq master mix

Appendix 4.2 Primer sequences and annealing temperatures for custom amplicon sequencing of *TNFSF9*

TNFSF9 Fragment	Primer Sequences 5' to 3'	Product size	PCR reaction	Cycling conditions
Amplicon 1	F: tcttaatttcaacattcacattccaggt R: agtcataaggtagacttctcattcaaca	2802 bp	5 µL Phusion Taq (NEB Inc.), 0.8 µL each of F and R primers at 10 µM, 2.4 µL H_2O, 1 µL DNA at 25 ng/µL	66ºC annealing temperature, 40 cycles, 2 min extension time
Amplicon 2	F: tctgttgaatgagaagtctaccttatga R: aatgtcctaagtcctacatagagaaagg	3540 bp	5 µL GoTaq Green, 0.8 µL each of F and R primers at 10 µM, 0.5 µL 100% DMSO, 1.9 µL H_2O, 1 µL DNA at 25 ng/µL	55ºC annealing temperature, 45 cycles, 3 min 30 sec extension time
Amplicon 3	F: cctttctctatgtaggacttaggacatt R: cctggcctcagtgtgaagat	3015 bp	5 µL MyTaq HS, 0.8 µL each of F and R primers at 10 µM, 2.4 µL H_2O, 1 µL DNA at 25 ng/µL	95ºC 1 min 95ºC 15 sec 63ºC 15 sec 72ºC 2min (45 cycles) 72ºC 5 min 4ºC hold
Amplicon 4	F: ctgacatgttcggtgctcag R: attcctcacgtattaccataactcagat	1084 bp	5 µL MyTaq HS, 0.8 µL each of F and R primers at 10 µM, 2.4 µL H_2O, 1 µL DNA at 25 ng/µL	95ºC 1 min 95ºC 15 sec 65ºC 15 sec 72ºC 45 sec (40 cycles) 72ºC 5 min 4ºC hold
Amplicon 5	F: atctgagttatggtaatacgtgaggaat R: ggaaaagaaacctcaaacaagaaaaaca	3629 bp	5 µL Phusion Taq (NEB Inc.), 0.8 µL each of F and R primers at 10 µM, 2.4 µL H_2O, 1 µL DNA at 25 ng/µL	66ºC annealing temperature, 40 cycles, 2 min extension time

Appendix 4.3 Samples selected for custom amplicon sequencing of *TNFSF9*

Individual ID	Sex	TFHMS case type	HM Subtype	Age at Diagnosis
LK0001-001	M	Familial	MALT	72
LK0002-099	F	Familial	NHL	39
LK0004-012	M	Familial	HL	23
LK0016-001	M	Familial	CLL	76
LK0016-108	M	Familial	MM	72
LK0016-137	M	Familial	CLL	63
LK0016-187	F	Familial	FL	56
LK0024-001	M	Familial	HCL	59
LK0051-001	M	Familial	T-cell NHL	31
LK0124-117	M	Familial	AML	24
LK0153-065	M	Familial	MDS to AML RAEB-2	63
LK0508-001	F	Non-familial	AML	18
LK0509-001	F	Non-familial	PV	31
LK0537-001	F	Familial	CLL	66
LK0537-002	M	Familial	CLL	58
LK0580-004	M	Familial	NHL	39
LK0585-001	F	Non-familial	MCL	64
LK0595-001	M	Non-familial	CLL	70
LK0625-001	F	Familial	MM	52
LK0625-003	M	Familial	CLL	60
LK0627-001	F	Non-familial	B cell gastric NHL	74
LK0628-001	F	Non-familial	DLBCL	68
LK0629-001	M	Non-familial	AML	40
LK0633-001	M	Non-familial	DLBCL	44
LK0634-001	M	Non-familial	HL	18
LK0635-001	M	Non-familial	DLBCL	57
LK0647-001	M	Familial	CLL	68
LK0650-001	M	Non-familial	CLL	58
LK0654-001	F	Non-familial	CML	29
LK0658-001	M	Non-familial	CLL	76
LK0660-001	M	Non-familial	WM	33
LK0661-001	M	Non-familial	NHL	69
LK0664-001	F	Non-familial	FL	41
LK0672-001	F	Familial	ALL	18
LK0673-001	F	Non-familial	NHL	33
LK0687-001	M	Non-familial	CLL	67
LK0689-001	F	Non-familial	HL	66
LK0692-001	M	Non-familial	CLL	53
LK0693-001	F	Non-familial	ET	70
LK0705-001	F	Non-familial	NHL	33

Individual ID	Sex	TFHMS case type	HM Subtype	Age at Diagnosis
LK0717-001	M	Non-familial	NHL	47
LK0732-001	M	Non-familial	NHL	52
LK0737-001	M	Non-familial	HCL	53
LK0739-001	F	Non-familial	NHL	80
LK0756-001	M	Non-familial	DLBCL	62
LK0760-001	F	Non-familial	DLBCL	65
LK0786-001	M	Familial	CLL	74
LK0788-001	F	Non-familial	NHL	86
LK0802-001	M	Non-familial	HCL	70
LK0804-001	F	Non-familial	T-cell NHL	31
LK0811-001	M	Non-familial	DLBCL	61
LK0818-001	M	Non-familial	NHL	59
LK0823-001	M	Familial	FL	51
LK0823-002	F	Familial	NHL	54
LK0836-001	M	Familial	CLL	65
LK0836-002	M	Familial	HCL	53
LK0848-001	M	Non-familial	NHL	64
LK0849-001	F	Non-familial	NHL	56
LK0853-001	F	Non-familial	NHL	Unknown
LK0855-001	F	Non-familial	HL	51
LK0927-001	M	Non-familial	WM	57
LK0933-001	M	Non-familial	HCL	67
LK0956-001	M	Non-familial	DLBCL	47
LK0958-001	M	Non-familial	HCL	49
LK0981-001	M	Non-familial	DLBCL	58
LK0984-001	M	Non-familial	HL	13
LK1033-001	F	Non-familial	BL	67
LK1040-001	M	Non-familial	FL	55
LK1045-001	F	Non-familial	HL	35
LK1155-001	F	Familial	CLL	60
LK2447-002	F	Familial	FL	60
LK6000-031	F	Familial	NHL	38
LK6000-141	F	Familial	DLBCL	42
LK7738-001	M	Non-familial	NHL	Unknown
LK7739-001	F	Familial	AML	78
LK7740-001	F	Familial	MM->CML (BCR-ABL +ve)	Unknown
LK7740-002	M	Familial	CLL->CML (BCR-ABL +ve)	Unknown
LK7743-001	F	Familial	CLL	48
LK7743-002	F	Familial	CLL	77
LK7746-001	M	Non-familial	FL	Unknown
LK7748-001	M	Familial	MCL	62
LK7749-001	F	Familial	MM	73
LK7750-002	M	Familial	MDS	70
LK7751-001	F	Familial	FL	75
LK7752-001	F	Non-familial	FL	53

Individual ID	Sex	TFHMS case type	HM Subtype	Age at Diagnosis
LK7753-001	M	Non-familial	NHL	56
LK7754-001	F	Familial	ET	80
LK7755-001	M	Familial	AML	79
LK7755-004	M	Familial	NHL	50
LK7756-004	M	Familial	AML	Unknown
LK7759-001	F	Non-familial	B Cell	Unknown
LK7766-001	F	Familial	HL	22
LK7768-001	F	Familial	CLL	72
LK7772-001	F	Familial	NHL	66
LK7773-001	F	Familial	MM	73
LK7773-002	M	Familial	NHL	73

Appendix 4.4 Representative TaqMan genotyping results for *TDP2* rs200729372 and *TNFSF9* rs61750000

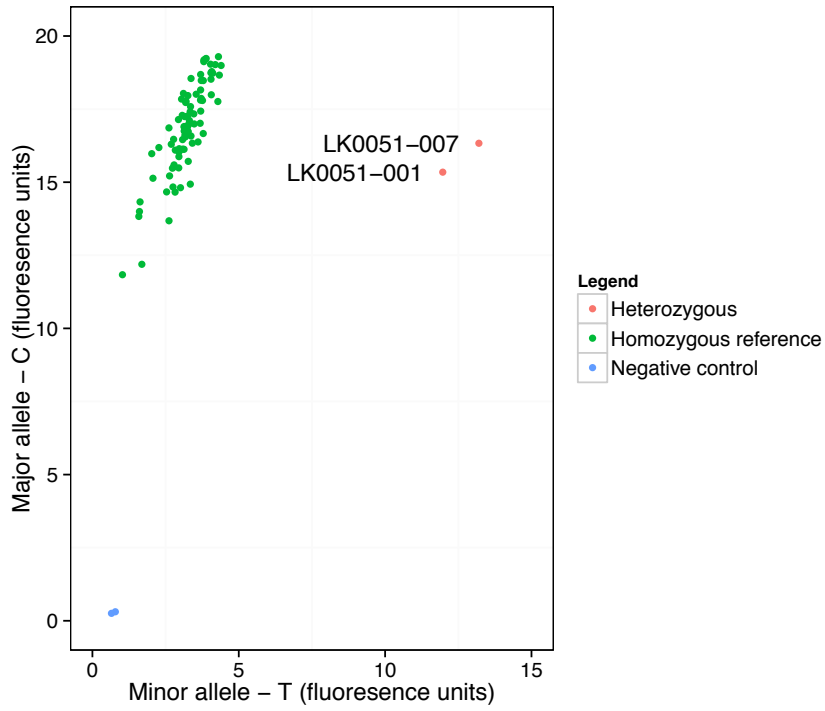


Figure 1. Representative TaqMan genotyping results from 94 samples for *TDP2* rs200729372, including two known heterozygous samples.

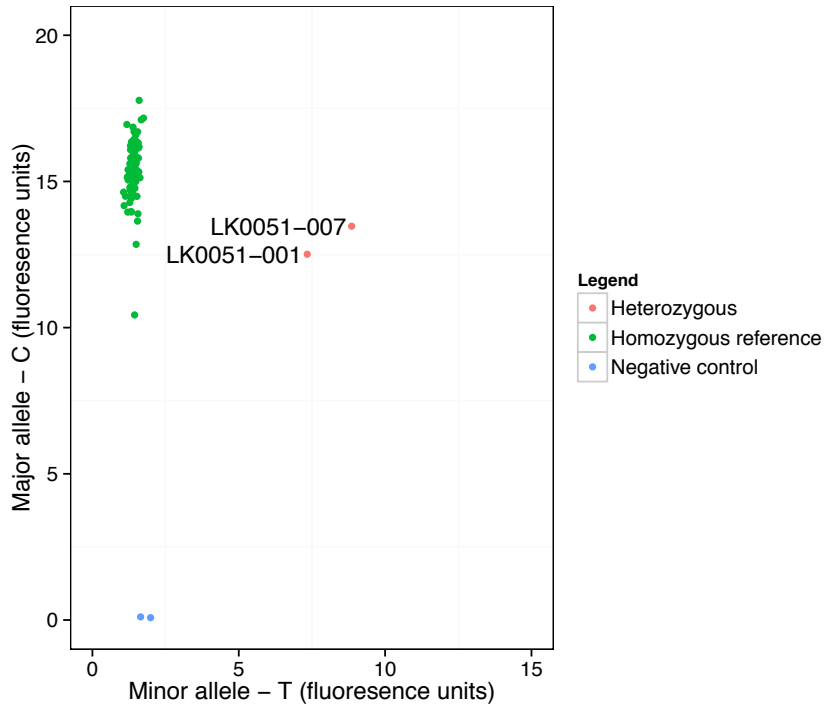


Figure 2. Representative TaqMan genotyping results from 94 samples for *TNFSF9* rs61750000, including two known heterozygous samples.

Appendix 4.5 Microarray gene expression profiles for *TDP2* and *TNFSF9*

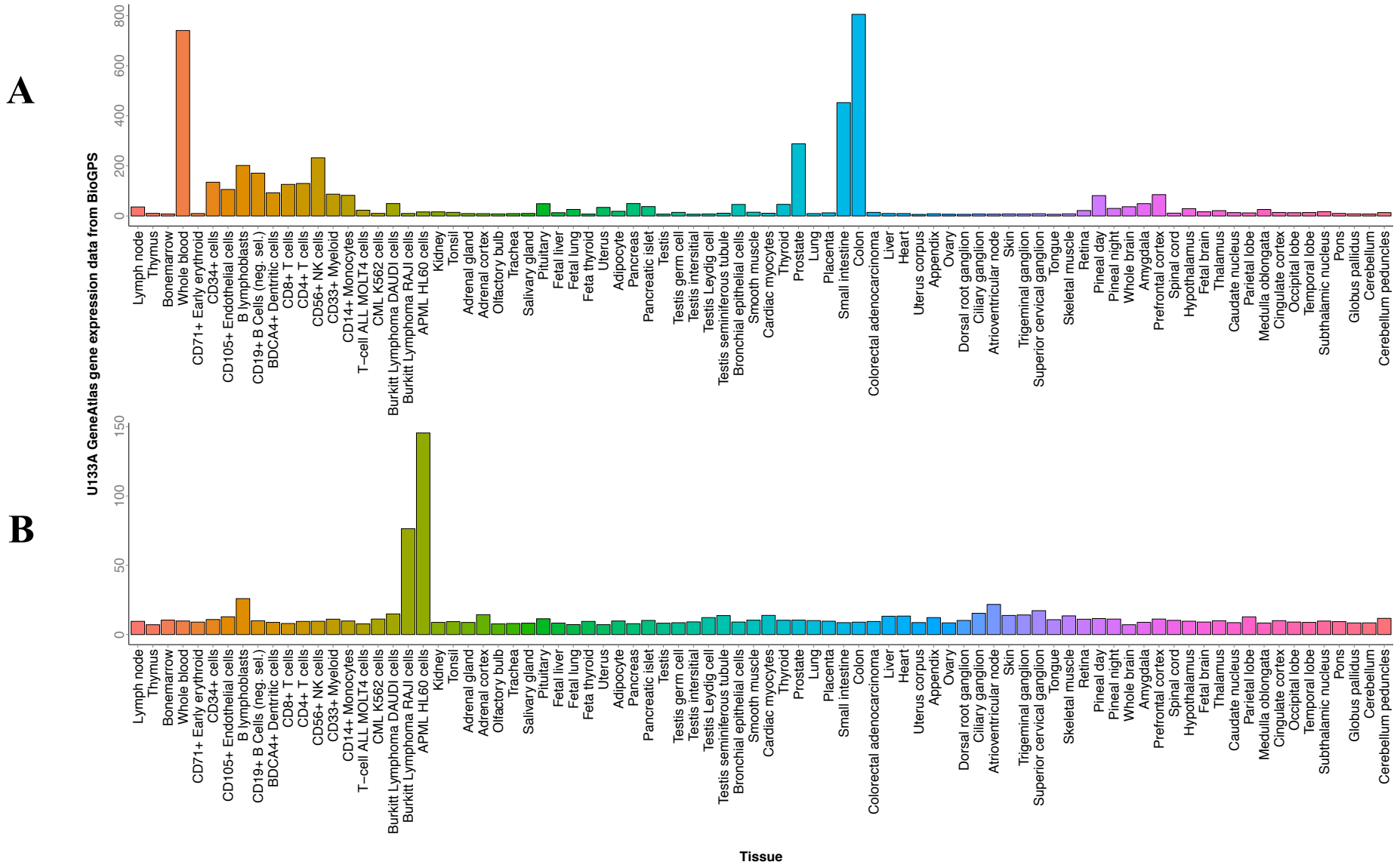


Figure 1 BioGPS sourced U133A GeneAtlas microarray gene expression profiles[249, 250] for *TDP2* (A) and *TNFSF9* (B).

Appendix 4.6 *TNFSF9* G139A sequence logo and protein structure

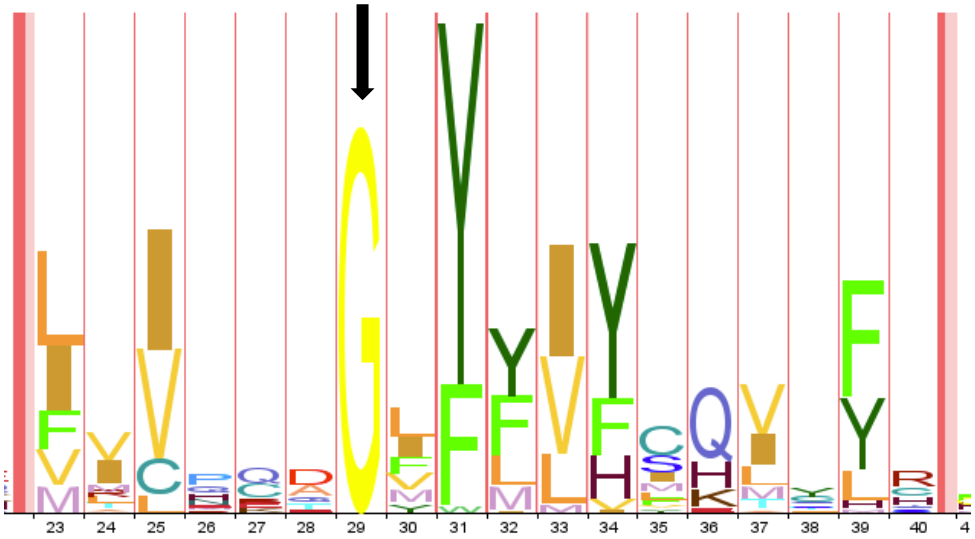


Figure 1. Partial protein alignment sequence logo for the TNF superfamily protein domain with the *TNFSF9* G139A residue indicated by the arrow.

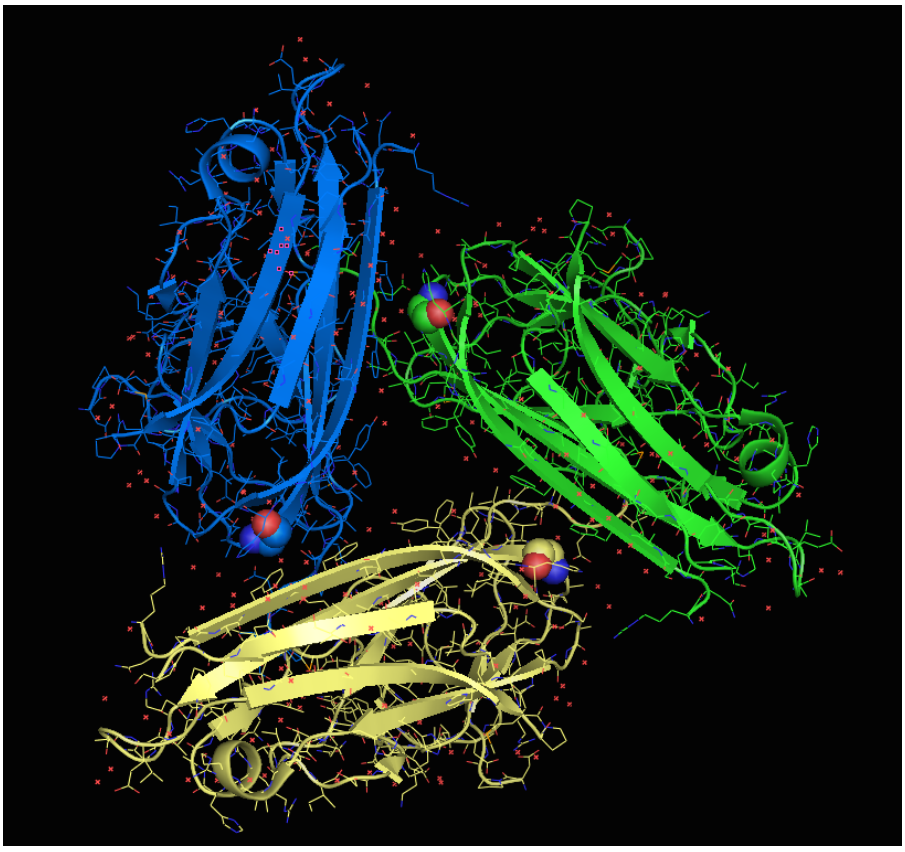


Figure 2. TNFSF9 homotrimer protein structure of the TNF superfamily protein domain with individual domains indicated by colour and the altered G139A amino acid residue indicated by a molecular ball.

Appendix 5.1 Chapter 5 Publication

The following article was published in *Oncology Reports*.

Blackburn, N. B. *et al.* A retrospective examination of mean relative telomere length in the Tasmanian Familial Hematological Malignancies Study. *Oncol Rep* 33, 25–32 (2015).

A retrospective examination of mean relative telomere length in the Tasmanian Familial Hematological Malignancies Study

NICHOLAS B. BLACKBURN[1*], JAC C. CHARLESWORTH[1,2*], JAMES R. MARTHICK[1], ELIZABETH M. TEGG[3,4], KATHERINE A. MARSDEN[1,3], VELANDAI SRIKANTH[5], JOHN BLANGERO[2], RAY M. LOWENTHAL[1,3], SIMON J. FOOTE[6] and JOANNE L. DICKINSON[1]

[1]Menzies Research Institute Tasmania, University of Tasmania, Hobart, TAS 7000, Australia; [2]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX 78245-0549, USA; [3]Royal Hobart Hospital, Hobart, TAS 7001; [4]School of Medicine, University of Tasmania, Hobart, TAS 7000; [5]Department of Medicine, Monash Medical Centre, Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, VIC 3168; [6]John Curtain School of Medical Research, Australian National University, ACT 2601, Australia

Received September 3, 2014; Accepted October 2, 2014

DOI: 10.3892/or.2014.3568

Abstract. Telomere length has a biological link to cancer, with excessive telomere shortening leading to genetic instability and resultant malignant transformation. Telomere length is heritable and genetic variants determining telomere length have been identified. Telomere biology has been implicated in the development of hematological malignancies (HMs), therefore, closer examination of telomere length in HMs may provide further insight into genetic etiology of disease development and support for telomere length as a prognostic factor in HMs. We retrospectively examined mean relative telomere length in the Tasmanian Familial Hematological Malignancies Study using a quantitative PCR method on genomic DNA from peripheral blood samples. Fifty-five familial HM cases, 191 unaffected relatives of familial HM cases and 75 non-familial HM cases were compared with 758 population controls. Variance components modeling was employed to identify factors influencing variation in telomere length. Overall, HM cases had shorter mean relative telomere length ($P=2.9\times10^{-6}$) and this was observed across both familial and non-familial HM cases ($P=2.2\times10^{-4}$ and 2.2×10^{-5}, respectively) as well as additional subgroupings of HM cases according to broad subtypes. Mean relative telomere length was also significantly heritable (62.6%; $P=4.7\times10^{-5}$) in the HM families in the present study. We present new evidence of significantly shorter mean relative telomere length in both familial and non-familial HM cases from the same population adding further support to the potential use of telomere length as a prognostic factor in HMs. Whether telomere shortening is the cause of or the result of HMs is yet to be determined, but as telomere length was found to be highly heritable in our HM families this suggests that genetics driving the variation in telomere length is related to HM disease risk.

Correspondence to: Dr Joanne L. Dickinson, Menzies Research Institute Tasmania, University of Tasmania, Private Bag 23, 17 Liverpool Street, Hobart, TAS 7000, Australia
E-mail: jo.dickinson@utas.edu.au

*Contributed equally

Key words: telomere length, hematological malignancies, familial cancer

Introduction

Telomeres are DNA-protein structures at chromosome ends consisting of repeating hexameric nucleotide sequences of TTAGGG (1). The primary role of the telomere is to cap chromosome ends to prevent aberrant recombination as a result of exposed chromosomal DNA; making telomeres essential for maintenance of genomic integrity (2). With each cell division telomeres shorten due to the incomplete DNA replication of chromosome ends by DNA polymerases eventually triggering cell senescence or apoptosis to prevent further shortening and exposure of chromosomal DNA (3). The telomerase complex can counteract telomere shortening in actively dividing cells by catalyzing the addition of TTAGGG repeats to chromosome ends (3,4), however, this is not a full restoration and telomeres progressively shorten with age (3,5,6). A number of studies have reported an association between telomere length in lymphocytes and an increased risk of age-related diseases including cancer (7,8). To date, the main understanding of the role of telomeres in cancer is that excessive telomere shortening leads to increased genetic instability and chromosomal end-to-end fusions (2,9), which then leads to a malignant cell transformation (9).

Studies of monozygotic and dizygotic twins and large families have established a genetic component to the determination of telomere length. Estimates of the heritability of telomere length ranges between 78 and 82% in studies of twins and sibling pairs (which generally produce inflated heritability estimates) (10,11) and 44% in a study of large Amish families (12). Although it has been proposed that the heritability of

telomere length can be accounted for by shared environmental factors (13) the consensus is that telomere length is primarily determined by parental inheritance including at least partial inheritance of chromosome-specific telomere lengths (14,15). This view is strongly supported by mouse models of telomere length inheritance (16).

Telomere length is a proposed risk factor for cancer given its importance in maintaining genomic integrity (2,9) and shorter telomeres have been shown to be associated with a range of cancers (7,8,17). The proposed role of telomere length as a cancer risk factor, coupled with its high heritability, raises the idea that telomere length may also be important in familial cancers such as familial hematological malignancies (HMs). Additionally, inherited changes in telomere length as a result of mutations in the telomere maintenance genes *TERT* and *TERC* have been identified in four HM families with myelodysplastic syndrome and acute myeloid leukemia (MDS-AML) (18). Therefore, telomere length may be a risk factor for other HM subtypes.

Hematological malignancies arise as a result of the neoplastic transformation of cells involved in hematopoiesis and include a broad range of subtypes of leukemias, lymphomas and myelomas (19). A well-established risk factor for HMs is family history indicating that familial HMs have an underlying genetic component. For example, a population-based study of the Swedish Cancer Registry showed an 8.5-fold increase in the risk of developing chronic lymphocytic leukemia (CLL) in first-degree relatives of CLL patients as well as an increased risk of developing other HM subtypes providing evidence for a shared genetic etiology across HM subtypes (20).

There is accumulating evidence that implicates telomere length as an important factor in the development of a range of HMs. This includes studies of several different HM subtypes where shorter telomeres were found in circulating tumor cells (21-26). While these studies have revealed important insights into telomere dynamics in circulating tumor cells there has been less focus on prospective and retrospective studies of pre-disease and remission telomere lengths respectively in HM patients. Indeed, one prospective study of telomere length in pre-disease blood samples surprisingly showed that longer telomere length was associated with a future risk of NHL (non-Hodgkin lymphoma) (27). Furthermore, in a study of chronic leukemia, Mansouri and colleagues (28) elegantly showed that telomere length has potential as a clinical prognostic marker in HMs. In their study, patients with shorter telomeres were associated with high-risk genetic markers and in patients with otherwise good prognostic markers, telomere length was an independent prognostic factor that subdivided the good prognosis group into groups with distinct outcomes. Therefore, there is potential for telomere length to be a clinically relevant prognostic risk factor for HMs.

The aim of the present study was to explore whether telomere length is involved in familial HMs and to find new evidence supporting telomere length as a prognostic risk factor for HMs. To this end, we examined telomere length in the Tasmanian Familial Hematological Malignancies Study (TFHMS); a genetic study comprising large Tasmanian families with multiple cases of HMs and a collection of population matched non-familial HM cases (29-31) and controls (32,33). Previously, similar collections have focused on families with one predominant subtype of HM. The TFHMS includes families with a dense aggregation of several HM subtypes across multiple generations; for example family LK2042 includes 32 cases in five generations (Table I).

The strength of a familial approach to examining telomere length lies in the enrichment of the shared genetic backgrounds between related individuals as related cases are likely to share genetic variants contributing to variation in telomere length which may in turn be affecting their risk of developing HMs. In the present study, we used the TFHMS to measure the heritability of telomere length as a quantitative trait in the study families and then examine whether HMs account for measured variation in telomere length.

Materials and methods

Ethics statement. The TFHM study was approved by the Human Research Ethics Committee (Tasmanian network), reference number: H8551, and written informed consent was obtained from all participating individuals.

The Tasmanian Familial Hematological Malignancies Study. As previously described (29) during the period 1972-1980 all patients with HMs diagnosed in Tasmania (the island state of Australia) were invited to participate in a population-based study examining the association of occupation and place of residence with risk of development of myeloproliferative and lymphoproliferative disorders (34,35). Using a genealogical database at the Menzies Research Institute Tasmania the individuals participating in the original population-based study were linked to both current generations and records from the Tasmanian Cancer Registry, which has documented cases of HMs since 1978. Family members provided further information through questionnaires and personal interviews. This allowed us to form pedigrees of Tasmanian families with multiple cases of HMs as well as a collection of HM cases with no reported family history of disease (Table I).

Confirmation of diagnosis was, where possible, obtained for all cases and in particular 13 study families were classified by a single experienced hematologist (E.M.T.) according to the 2008 World Health Organization classification (19) as previously described (29). For the remaining study families, case diagnosis was obtained from the Tasmanian Cancer Registry records and by review of available pathology reports of cases that consented to participate in the TFHMS. More extensive clinical information is not currently available due to the multi-center and multi-specialist nature of the original data collection.

Study samples. In this TFHM-based study we used DNA obtained from peripheral blood samples from 55 familial HM cases, 191 unaffected relatives of familial cases and 75 non-familial cases. DNA from 40 TFHMS families was available for the present study with samples available from both HM cases and unaffected relatives in 14 families. The remaining families were comprised of samples from HM cases with a known family history of disease alone or from unaffected relatives of HM cases. Of the 191 unaffected relatives, 171 were first-degree relatives of HM cases and the remaining subjects were more distantly related or spouses. For HM cases,

Table I. Summary of the TFHMS families used in this study.

Family	Known HM cases	Generations with HM cases	HM cases with telomere length measurement	Unaffected relatives with telomere length measurement
LK0001	14	4	1	16
LK0002	15	3	1	5
LK0004	7	2	1	11
LK0016	18	5	2	19
LK0024	3	2	1	0
LK0026	6	2	1	5
LK0040	7	4	2	2
LK0051	21	5	3	26
LK0054	9	3	0	2
LK0065	8	2	0	8
LK0124	24	5	2	34
LK0132	5	2	0	7
LK0139	7	2	1	2
LK0153	9	2	3	2
LK0511	2	2	1	0
LK0512	2	1	1	0
LK0537	2	1	2	0
LK0546	2	2	1	0
LK0560	2	2	1	0
LK0561	2	2	1	0
LK0600	5	3	2	0
LK0625	4	2	2	0
LK0647	2	2	1	0
LK0672	3	3	1	0
LK0836	6	3	2	5
LK1155	2	1	1	3
LK2042	32	5	6	40
LK2447	3	2	1	2
LK6000	6	2	1	0
LK7739	2	1	1	0
LK7740	2	2	2	0
LK7743	3	2	2	0
LK7744	2	2	0	1
LK7748	2	2	1	0
LK7749	3	2	1	0
LK7750	4	2	2	0
LK7751	9	3	1	0
LK7754	3	1	1	0
LK7755	2	2	1	0
LK7768	2	1	1	0
Non-familial cases	-	-	75	1

HM, hematological malignancy; TFHM, Tasmanian Familial Hematological Malignancies Study.

DNA was collected from 1 month to 64.9 years post HM diagnosis (mean, 9.9 years). Population controls were recruited randomly from the Tasmanian electoral role (n=758) through the TASCOG study (33) (a population-based study of gait in older Tasmanians) or obtained from the control samples in a Tasmanian familial prostate cancer case-control study (32) both conducted at the Menzies Research Institute Tasmania. Details concerning the participants in this study are shown in Table II, and the distribution of HM case subtypes in this study is summarized in Table III. Non-familial HM cases had no self-reported family history of HMs and did not appear in any of our study families after thorough genealogical examination. Frequent updates from the Tasmanian Cancer Registry were used to monitor the occurrence of HMs in the unaffected relatives that are part of this study. Population control DNA samples were also extracted from peripheral blood samples in the same laboratory, using the same methodology as the TFHMS samples. Genomic DNA was extracted from peripheral blood samples using the Nucleon BACC 3 DNA Extraction kit (GE Healthcare).

Telomere length measurement. We investigated the mean relative telomere length in peripheral blood samples using a slightly amended protocol for a validated monochrome multiplex quantitative PCR method outlined by Cawthon (36). This method measures the relative telomere length by calculating the ratio, T/S, between telomere repeat copy number amplification (T) and the amplification of a single-copy gene, albumin (S). The average T/S ratio was obtained as the mean of the triplicate measurements for each sample. Individual measurements were excluded from the average T/S ratio calculation when the replicate failed or a large standard error was observed. The coefficient of variation calculated across all assay plates using repeated cross-plate samples was 3.4%.

Telomere length measurement was performed in 10 μl volumes using a LightCycler 480 in a 96-well plate format. Each 96-well plate contained a six point standard curve 2, 5, 15, 50, 100 and 150 ng, a unique sample common to each plate, a no template control and 24 unknown case/control samples all repeated in triplicate, with 1.6% sample replication across plates. The genomic DNA used for the standard curve was from a 27-year-old female control study participant.

Final reagent concentrations were 5 ng of genomic DNA, primer telg 200 nM (5'-ACACTAAGGTTTGGGTTTGGGTTTGGGTTTGGGTTAGTGT-3'), primer telc 700 nM (5'-TGTTAGGTATCCCTATCCCTATCCCTATCCCTATCCCTAACA-3'), primer albu 500 nM (5'-CGGCGGCGGGCGGCGCGGGCTGGGCGGAAATGCTGCACAGAATCCTTG-3'), primer albd 500 nM (5'-GCCCGGCCCGCCGCGCCCGTCCCGCCGGAAAAGCATGGTCGCCTGTT-3'), AmpliTaq Gold (Applied Biosystems) 0.625 U, GeneAmp 10X PCR buffer (Life Technologies) containing 50 mM KCl, 10 mM Tris-HCl pH 8.3 and 1.5 mM $MgCl_2$, 1 mM DTT, 1 M Betaine (Sigma-Aldrich), 0.0025 mM Syto9 (Life Technologies) and 0.25 mM of each dNTP (Bioline). Cycling conditions were as follows: 95˚C for 15 min, 2 cycles of 94˚C for 15 sec, 49˚C for 60 sec; four cycles of 84˚C for 20 sec, 59˚C for 30 sec, then 40 cycles of 94˚C for 15 sec, 59˚C for 30 sec with signal acquisition for telomere repeat copy number amplification, 84˚C for 30 sec, then 85˚C for 20 sec with signal acquisition for albumin amplification. A melting curve was generated for each plate. Ct values were calculated using LinRegPCR (37) and a standard

Table II. Mean age, sex distribution and relative telomere length in the sample groups.

Sample group	N	Male sex, n (%)	Mean age (range)	Mean relative T/S ratio[a] (95% CI)
Controls	758	578 (76.3)	67.51 (30.67-87.97)	0.64 (0.62-0.66)
Unaffected relatives of HM cases	191	77 (40.3)	61.65 (27.26-92.95)	0.73 (0.69-0.76)
All HM cases	130	73 (56.2)	65.14 (13.24-95.53)	0.53 (0.50-0.56)
Familial HM cases	55	32 (58.2)	64.45 (13.24-87.45)	0.57 (0.52-0.63)
Non-familial HM cases	75	41 (54.7)	68.79 (22.42-95.53)	0.50 (0.46-0.53)

[a]Mean relative T/S ratio is the ratio between telomere repeat copy number (T) and a single-copy gene, *ALB*, copy number (S), a measure of mean relative telomere length. CI, confidence interval; HM, hematological malignancy. Mean age (range) is expressed in years.

Table III. Disease characteristics of study samples.

	HM familial cases, n (%)	HM non-familial cases, n (%)	All HM cases, n (%)
HM subtypes			
Acute lymphoblastic leukemia	2 (3.6)	0	2 (1.5)
Acute myeloid leukemia	5 (9.1)	8 (10.7)	13 (10.0)
Chronic myeloid leukemia	0	3 (4.0)	3 (2.3)
Essential thrombocythemia	1 (1.8)	1 (1.3)	2 (1.5)
Hodgkin lymphoma	5 (9.1)	4 (5.3)	9 (6.9)
Myelodysplastic syndrome	2 (3.6)	0	2 (1.5)
Myeloproliferative neoplasm	1 (1.8)	2 (2.7)	3 (2.3)
T-cell non-Hodgkin lymphoma	1 (1.8)	2 (2.7)	3 (2.3)
Mature B cell neoplasms			
Non-Hodgkin lymphoma unclassified	2 (3.6)	10 (13.3)	12 (9.2)
Chronic lymphocytic leukemia	12 (21.8)	12 (16.0)	24 (18.5)
Diffuse large B-cell lymphoma	4 (7.3)	10 (13.3)	14 (10.8)
Follicular lymphoma	4 (7.3)	9 (12.0)	13 (10.0)
Multiple myeloma	7 (12.7)	5 (6.7)	12 (9.2)
Other[a]	9 (16.4)	9 (12.0)	18 (13.8)
Total	55	75	130

[a]Other includes Burkitt lymphoma, hairy cell leukemia, lymphoma of mucosa-associated lymphoid tissue and Waldenström macroglobulinemia. HM, hematological malignancy.

curve was generated for both the telomere and albumin PCRs. A linear regression of the standard curve measurement values was used to correct for any variation in fluorescence levels derived from small fluctuations in DNA concentration. The equations from the linear regression of each standard curve were then used to calculate the log(DNA) value for the unknown case/control samples.

Statistical analysis. Average T/S ratios greater than 4 standard deviations from the control mean were excluded as outliers. Mean relative T/S ratios with 95% confidence intervals (CI) are reported in Table II. For analysis mean relative T/S ratios were transformed to fit a normal distribution using the inverse-normalization option in SOLAR (version 6.6.2) (38,39) to prevent non-normal distribution errors. In order to fully utilize the extended pedigree study design, correct for relatedness, and to maximize the information provided by telomere length as a quantitative trait we used variance components modeling in SOLAR (38,39) to determine the heritability of telomere length (adjusting for kinship and significant covariates) and to calculate the association between telomere length and disease. The primary benefit to using SOLAR is its ability to incorporate relatedness through the use of a kinship matrix and to fully utilize the quantitative trait data, which increases the power and accuracy of the trait heritability calculation.

Sex, age, age^2 and their interactions were included as covariates in all relevant analyses. Potential batch effects were adjusted for by applying household modeling (38,39) by coding each assay plate as a separate household. SOLAR has been previously used in the analysis of telomere length in related

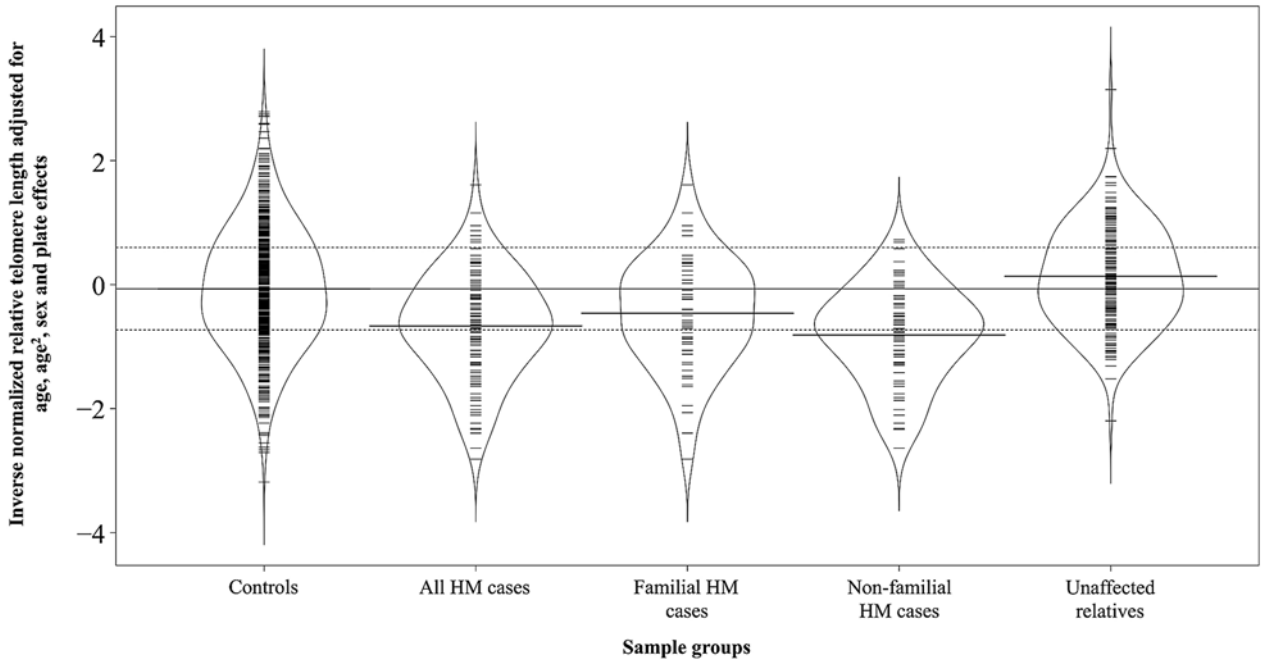


Figure 1. Bean plot quartile analysis of adjusted inverse normalized relative telomere lengths. The adjusted inverse normalized relative telomere length for each group is displayed as a bean plot with individual sample measurements as lines within the bean and the overall distribution of all samples in each group shown. Horizontal bars for each bean indicate the mean of each group. The solid line and dashed lines show the mean and interquartile range of the control group.

individuals (11,12,40). The algorithms utilized in SOLAR for the analysis of quantitative traits in related individuals are more appropriate to employ in the present study than a more traditional approach of analyzing quantitative traits using percentiles or quartiles. Nevertheless, we also present observations from a quartile analysis of inverse normalized relative T/S ratios adjusted for age, sex and batch effects in SOLAR with quartiles defined from the adjusted T/S ratios in the control population (Fig. 1). Bean plots in Fig. 1 were constructed using the R package 'beanplot' (41).

Results

Mean relative telomere length in familial and non-familial HM cases, unaffected relatives and control subjects was measured by monochrome multiplex quantitative PCR. Using SOLAR we found that the heritability of mean relative telomere length was 62.5% ($P=4.7x10^{-5}$, SE=0.14). The removal of HM cases (n=130) from analysis only marginally altered the heritability of mean relative telomere length (75.5%; $P=1.2x10^{-5}$, SE=0.15).

The use of variance components modeling in SOLAR permits appropriate statistical analyses inclusive of familial relationships. These analyses revealed that disease status was significantly associated with mean relative telomere length (Table IV, primary analysis model 1; $P=2.9x10^{-6}$) with HM cases having shorter mean relative telomere length when compared with unaffected individuals. We conducted a separate analysis distinguishing familial and non-familial cases. Familial cases and non-familial cases each had significantly shorter mean relative telomere length (Table IV; primary analysis model 2, $P=2.2x10^{-4}$ and P=2.2x10-5, respectively).

The most frequent type of HM diagnosed in the present study was mature B cell neoplasms (MBCNs; Table III). Analysis of MBCNs as one group and HMs other than MBCNs as a second group (combined due to the small numbers of other subtypes) showed that both groupings had shorter mean relative telomere length than unaffected individuals from both study families and population controls (Table IV; primary analysis model 3, $P=3.5x10^{-5}$ and $P=9.3x10^{-5}$, respectively). These groupings were then divided according to whether the HM case was familial or non-familial. Analysis showed that all HM case subgroupings maintained significantly shorter mean relative telomere length (Table IV; primary analysis model 4). An analysis using specific HM subtypes was not possible due to not having enough statistical power at this level of HM classification with small numbers of HM subtypes (Table III).

Variance components modeling also identified age and sex (Table IV; primary analysis model 1, $P=4.8x10^{-8}$ and $P=4.0x10^{-3}$, respectively) as significant covariates for mean relative telomere length variation across all models. Mean relative telomere length declined with age and males had shorter telomeres than females. Age^2 was also a significant covariate, indicating that the decline in mean relative telomere length with age has a non-linear component (Table III; primary analysis model 1, $P=8.0x10^{-3}$).

Four sub-analyses of the primary data were also performed to determine whether particular features of the study population were contributing to the disease associations found in the primary analysis models (Table IV). Sub-analyses included exclusion of HM cases, controls and unaffected relatives 80 years or older (n=126), exclusion of CLL cases (n=24), exclusion of cases with samples collected within two years of diagnosis (n=24) as well as all three exclusions together (n=162, some individuals were in multiple exclusion categories). In each sub-analysis the principle findings from the primary analysis models were maintained.

Table IV. Variance component modeling analysis of inverse normalized mean relative telomere length primary analysis and sub-analyses with exclusions.

Models and variables	Primary analysis P-values	≥80 years old excluded (n=126) P-values	CLL cases excluded (n=24) P-values	Possible malignant samples excluded[a] (n=24) P-values	All exclusions applied (n=162) P-values
Model 1					
Age	$4.8x10^{-8}$	$7.5x10^{-5}$	$1.6x10^{-8}$	$6.9x10^{-8}$	$3.4x10^{-5}$
Age^2	$8.0x10^{-3}$	0.04	$6.0x10^{-3}$	0.01	0.07
Sex	$4.0x10^{-3}$	0.01	$7.0x10^{-3}$	$2.0x10^{-3}$	0.01
All HM cases	$2.9x10^{-6}$	$7.3x10^{-6}$	$2.9x10^{-7}$	$1.1x10^{-4}$	$4.6x10^{-5}$
% trait variance accounted for by model	10.07%	9.46%	10.38%	9.56%	9.38%
Model 2					
Age	$4.3x10^{-8}$	$5.1x10^{-5}$	$1.6x10^{-8}$	$7.4x10^{-8}$	$3.7x10^{-5}$
Age^2	$8.0x10^{-3}$	0.04	$6.0x10^{-3}$	0.01	0.07
Sex	$3.0x10^{-3}$	0.01	$6.0x10^{-3}$	$2.0x10^{-3}$	$9.0x10^{-3}$
Familial HM cases	$2.2x10^{-4}$	$1.0x10^{-3}$	$1.6x10^{-5}$	0.01	$8.0x10^{-3}$
Non-familial HM cases	$2.2x10^{-5}$	$6.9x10^{-6}$	$7.1x10^{-5}$	$3.3x10^{-5}$	$2.7x10^{-5}$
% trait variance accounted for by model	10.62%	10.29%	10.55%	10.48%	10.00%
Model 3					
Age	$4.7x10^{-8}$	$7.2x10^{-5}$	$1.7x10^{-8}$	$7.7x10^{-8}$	$3.7x10^{-5}$
Age^2	$8.0x10^{-3}$	0.04	$6.0x10^{-3}$	0.01	0.07
Sex	$4.0x10^{-3}$	0.01	$6.0x10^{-3}$	$2.0x10^{-3}$	0.01
MBCNs	$3.5x10^{-5}$	$7.8x10^{-5}$	$5.5x10^{-6}$	$5.7x10^{-4}$	$3.8x10^{-4}$
HMs other than MBCNs	$9.3x10^{-5}$	$1.5x10^{-4}$	$3.4x10^{-5}$	$1.0x10^{-3}$	$5.8x10^{-4}$
% trait variance accounted for by model	10.08%	9.50%	10.37%	9.56%	9.39%
Model 4					
Age	$4.8x10^{-8}$	$4.1x10^{-5}$	$1.9x10^{-8}$	$6.6x10^{-8}$	$3.8x10^{-5}$
Age^2	$9.0x10^{-3}$	0.05	$7.0x10^{-3}$	0.02	0.08
Sex	$3.0x10^{-3}$	0.01	$6.0x10^{-3}$	$2.0x10^{-3}$	$9.0x10^{-3}$
Familial MBCNs	0.02	0.04	$3.0x10^{-3}$	0.18	0.07
Familial cases other than MBCNs	$5.2x10^{-4}$	$3.0x10^{-3}$	$5.7x10^{-4}$	0.01	0.04
Non-familial MBCNs	$2.4x10^{-5}$	$1.5x10^{-5}$	$1.3x10^{-4}$	$4.8x10^{-5}$	$1.5x10^{-4}$
Non-familial cases other than MBCNs	$2.0x10^{-3}$	$4.2x10^{-4}$	$3.0x10^{-3}$	$2.0x10^{-3}$	$3.2x10^{-4}$
% trait variance accounted for by model	10.87%	10.58%	10.57%	10.45%	10.05%

P-values for the significance of each trait or covariate were derived from variance component polygenic modeling in SOLAR. [a]HM case samples collected within ± 2 years of diagnosis were excluded. CLL, chronic lymphocytic leukemia; HM, hematological malignancy; MBCNs, mature B-cell neoplasms.

Categorization of cases into quartiles of mean relative telomere length determined from the distribution of age, sex and batch effect adjusted mean relative telomere length in controls (Fig. 1) showed that 43.1% of HM cases were in the lowest quartile of mean relative telomere length (below the lower interquartile dashed line), with 36.4% of familial HM cases and 48% of non-familial HM cases in the lowest quartile, whilst only 13.1% of unaffected relatives were in the lowest quartile. Similarly a low percentage of cases (5.4%) were in the longest quartile of mean relative telomere length (above the upper interquartile dashed line) whereas 28.3% of unaffected relatives were in the longest quartile. A clear trend for shorter mean relative telomere length in a higher percentage of HM cases was observed but this analysis did not permit familial relationships to be included in the analysis.

Discussion

These analyses determined that mean relative telomere length is highly heritable within the TFHMS families supporting previously reported heritability estimates in non-disease families (10-12). Our finding that mean relative telomere length was shorter in both familial and non-familial HM cases indicates that telomere length is likely to be important

in the genetic etiology of HMs. A previous study of mean relative telomere length in familial myelodysplastic syndrome MDS-AML has shown that affected individuals from four small families had shorter telomeres concurrent with mutations in the telomerase gene *TERT* and its RNA component *TERC* (18). Of the five cases across the four families reported to have shorter telomeres, two had aplastic anemia, two had MDS and one had MDS-AML. The present study extends the findings of Kirwan and colleagues (18) to that of large families with multiple HM subtypes finding new evidence of the involvement of telomere length in both familial and non-familial HMs.

Age, sex and age^2 as covariates explained a proportion of the variation in mean relative telomere length in the present study. This is in keeping with telomere length declining with age and males having shorter telomeres than females (7). A significant age^2 indicates a non-linear component in the age-related telomere length decline, a finding in line with a recent report showing a differential rate of decrease in telomere length over different age ranges (42). Our population controls did have a higher percentage of males, which could be suggested to be driving the association with sex, however, SOLAR was used to correct the mean relative telomere length for sex effects.

An important caveat with our retrospective study is that the finding of shorter mean relative telomere lengths in HM cases could also be related to disease susceptibility, treatment or the disease process. The present study did not have the necessary clinical information to appropriately analyze these factors. Currently, the literature surrounding the role of chemotherapeutic agents in telomere shortening remains controversial and inconclusive. Several studies in both HMs and other cancers such as breast cancer have shown that telomere length is unaffected when comparing pre- and post-chemotherapy measurements, when comparing patients who receive chemotherapy to those that do not or when comparing telomere length between patients and population controls (25,28,43,44). Other studies show a heterogeneous effect of chemotherapy on telomere length (45-47). It could be concluded from these reports and others that chemotherapy has no consistent influence on telomere length in blood cells particularly when examining multiple chemotherapeutic treatment regimens.

A second consideration is that shorter telomeres in HM cases could be the result of malignant cell DNA within the genomic DNA sample. We recognize that circulating malignant cells can be present for many years in chronic HM subtypes such as CLL. Based on the clinical diagnoses of HM cases in our study, we conducted two additional sub-analyses of the primary data. In the one analysis we removed all CLL cases (n=24) on the basis that DNA obtained from blood of cases with this subtype of HM was likely to contain DNA from diseased cells (Table IV). In the second analysis we removed all cases for which blood samples were obtained for DNA within 2 years of diagnosis (n=24; Table IV). Repeating the variance components modeling in these two analyses maintained the key significant associations with HM disease, suggesting that circulating disease did not contribute to the telomere length associations we have identified. In an additional sub-analysis we excluded all HM cases, controls and unaffected relatives (n=126) aged 80 years and above on the basis that the population HM risk increases with age. This did not change the principle findings of shorter telomeres in familial and non-familial HM cases nor did a final combined sub-analysis excluding individuals from all 3 sub-analyses. All cases, controls and unaffected relatives were included in the primary analysis models reported in Table IV.

In conclusion, our analyses showed for the first time that mean relative telomere length is heritable in large HM families with multiple generations affected by multiple subtypes of HMs, indicating a strong genetic effect driving trait variation. We also showed that both familial and non-familial HM cases from the same population had shorter mean relative telomere length. Taken together, the results from this retrospective study provide new evidence that mean relative telomere length is an important genetic factor in a wide range of HM subtypes and in individuals with and without a family history of disease. These findings contribute further support to the use of telomere length as a prognostic risk factor for HMs

Acknowledgements

The authors thank Annette Banks for her ongoing genealogical support. The authors would also like to thank the many years of dedicated research performed by Anne Piaszczyk, and the late Jean Panton. Finally we acknowledge the participants and their families in the Tasmanian Familial Hematological Malignancies study.

References

1. Moyzis RK, Buckingham JM, Cram LS, *et al*: A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. Proc Natl Acad Sci USA 85: 6622-6626, 1988.
2. Counter CM, Avilion AA, LeFeuvre CE, Stewart NG, Greider CW, Harley CB and Bacchetti S: Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. EMBO J 11: 1921-1929, 1992.
3. Harley CB, Futcher AB and Greider CW: Telomeres shorten during ageing of human fibroblasts. Nature 345: 458-460, 1990.
4. Greider CW and Blackburn EH: Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. Cell 43: 405-413, 1985.
5. Allsopp RC, Vaziri H, Patterson C, *et al*: Telomere length predicts replicative capacity of human fibroblasts. Proc Natl Acad Sci USA 89: 10114-10118, 1992.
6. Brouilette SW, Moore JS, McMahon AD, *et al*: Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study. Lancet 369: 107-114, 2007.
7. McGrath M, Wong JYY, Michaud D, Hunter DJ and de Vivo I: Telomere length, cigarette smoking, and bladder cancer risk in men and women. Cancer Epidemiol Biomarkers Prev 16: 815-819, 2007.
8. Willeit P, Willeit J, Mayr A, *et al*: Telomere length and risk of incident cancer and cancer mortality. JAMA 304: 69-75, 2010.
9. Artandi SE, Chang S, Lee SL, Alson S, Gottlieb GJ, Chin L and DePinho RA: Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. Nature 406: 641-645, 2000.
10. Slagboom PE, Droog S and Boomsma DI: Genetic determination of telomere size in humans: a twin study of three age groups. Am J Hum Genet 55: 876-882, 1994.
11. Vasa-Nicotera M, Brouilette S, Mangino M, *et al*: Mapping of a major locus that determines telomere length in humans. Am J Hum Genet 76: 147-151, 2005.
12. Njajou OT, Cawthon RM, Damcott CM, *et al*: Telomere length is paternally inherited and is associated with parental lifespan. Proc Natl Acad Sci USA 104: 12135-12139, 2007.

13. Huda N, Tanaka H, Herbert BS, Reed T and Gilley D: Shared environmental factors associated with telomere length maintenance in elderly male twins. Aging Cell 6: 709-713, 2007.
14. Graakjaer J, Bischoff C, Korsholm L, *et al*: The pattern of chromosome-specific variations in telomere length in humans is determined by inherited, telomere-near factors and is maintained throughout life. Mech Ageing Dev 124: 629-640, 2003.
15. Graakjaer J, Londoño-Vallejo JA, Christensen K and Kølvraa S: The pattern of chromosome-specific variations in telomere length in humans shows signs of heritability and is maintained through life. Ann NY Acad Sci 1067: 311-316, 2006.
16. Chiang YJ, Calado RT, Hathcock KS, Lansdorp PM, Young NS and Hodes RJ: Telomere length is inherited with resetting of the telomere set-point. Proc Natl Acad Sci USA 107: 10148-10153, 2010.
17. Wu X, Amos CI, Zhu Y, *et al*: Telomere dysfunction: a potential cancer predisposition factor. J Natl Cancer Inst 95: 1211-1218, 2003.
18. Kirwan MJ, Vulliamy T, Marrone A, *et al*: Defining the pathogenic role of telomerase mutations in myelodysplastic syndrome and acute myeloid leukemia. Hum Mutat 30: 1567-1573, 2009.
19. Swerdlow SH, Campo E, Harris NL, *et al*: WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. IARC, Lyon, 2008.
20. Goldin LR, Björkholm M, Kristinsson SY, Turesson I and Landgren O: Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. Haematologica 94: 647-653, 2009.
21. Wu K, Lund M, Bang K and Thestrup-Pedersen K: Telomerase activity and telomere length in lymphocytes from patients with cutaneous T-cell lymphoma. Cancer 86: 1056-1063, 1999.
22. Terasaki Y, Okumura H, Ohtake S and Nakao S: Accelerated telomere length shortening in granulocytes: a diagnostic marker for myeloproliferative diseases. Exp Hematol 30: 1399-1404, 2002.
23. Wu KD, Orme LM, Shaughnessy J, Jacobson J, Barlogie B and Moore MAS: Telomerase and telomere length in multiple myeloma: correlations with disease heterogeneity, cytogenetic status, and overall survival. Blood 101: 4982-4989, 2003.
24. Hartmann U, Brümmendorf TH, Balabanov S, Thiede C, Illme T and Schaich M: Telomere length and hTERT expression in patients with acute myeloid leukemia correlates with chromosomal abnormalities. Haematologica 90: 307-316, 2005.
25. Ghaffari SH, Shayan-Asl N, Jamialahmadi AH, Alimoghaddam K and Ghavamzadeh A: Telomerase activity and telomere length in patients with acute promyelocytic leukemia: indicative of proliferative activity, disease progression, and overall survival. Ann Oncol 19: 1927-1934, 2008.
26. Capraro V, Zane L, Poncet D, *et al*: Telomere deregulations possess cytogenetic, phenotype, and prognostic specificities in acute leukemias. Exp Hematol 39: 195-202.e2, 2011.
27. Lan Q, Cawthon R, Shen M, *et al*: A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of non-Hodgkin lymphoma. Clin Cancer Res 15: 7429-7433, 2009.
28. Mansouri L, Grabowski P, Degerman S, *et al*: Short telomere length is associated with NOTCH1/SF3B1/TP53 aberrations and poor outcome in newly diagnosed chronic lymphocytic leukemia patients. Am J Hematol 88: 647-651, 2013.
29. Tegg EM, Thomson RJ, Stankovich J, *et al*: Evidence for a common genetic aetiology in high-risk families with multiple haematological malignancy subtypes. Br J Haematol 150: 456-462, 2010.
30. Tegg EM, Thomson RJ, Stankovich JM, *et al*: Anticipation in familial hematologic malignancies. Blood 117: 1308-1310, 2011.
31. Lowenthal RM, Tegg EM and Dickinson JL: The Familial Tasmanian Haematological Malignancies Study (FaTHMS): its origins, its history and the phenomenon of anticipation. Transfus Apher Sci 49: 113-115, 2013.
32. FitzGerald LM, Patterson B, Thomson R, *et al*: Identification of a prostate cancer susceptibility gene on chromosome 5p13q12 associated with risk of both familial and sporadic disease. Eur J Hum Genet 17: 368-377, 2009.
33. Callisaya ML, Blizzard L, Schmidt MD, Martin KL, McGinley JL, Sanders LM and Srikanth VK: Gait, gait variability and the risk of multiple incident falls in older people: a population-based study. Age Ageing 40: 481-487, 2011.
34. Giles GG, Lickiss JN, Baikie MJ, Lowenthal RM and Panton J: Myeloproliferative and lymphoproliferative disorders in Tasmania, 1972-80: occupational and familial aspects. J Natl Cancer Inst 72: 1233-1240, 1984.
35. Lickiss JN, Giles GG, Baikie MJ, Lowenthal RM, Challis D and Panton J: Myeloproliferative and lymphoproliferative disorders in Tasmania, 1972-80: patterns in space and time. J Natl Cancer Inst 72: 1223-1231, 1984.
36. Cawthon RM: Telomere length measurement by a novel monochrome multiplex quantitative PCR method. Nucleic Acids Res 37: e21, 2009.
37. Ruijter JM, Ramakers C, Hoogaars WMH, Karlen Y, Bakker O, van den Hoff MJB and Moorman AFM: Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res 37: e45, 2009.
38. Almasy L and Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62: 1198-1211, 1998.
39. Almasy L and Blangero J: Variance component methods for analysis of complex phenotypes. Cold Spring Harb Protoc 2010: pdb.top77-pdb.top77, 2010.
40. Njajou OT, Blackburn EH, Pawlikowska L, *et al*: A common variant in the telomerase RNA component is associated with short telomere length. PLoS One 5: e13048, 2010.
41. Kampstra P: Beanplot: A boxplot alternative for visual comparison of distributions. J Statistical Software 28: 1-9, 2008. http://www.jstatsoft.org/v28/c01/.
42. Aubert G, Baerlocher GM, Vulto I, Poon SS and Lansdorp PM: Collapse of telomere homeostasis in hematopoietic cells caused by heterozygous mutations in telomerase genes. PLoS Genet 8: e1002696, 2012.
43. Schröder CP, Wisman GB, de Jong S, *et al*: Telomere length in breast cancer patients before and after chemotherapy with or without stem cell transplantation. Br J Cancer 84: 1348-1353, 2001.
44. Mirabello L, Garcia-Closas M, Cawthon R, *et al*: Leukocyte telomere length in a population-based case-control study of ovarian cancer: a pilot study. Cancer Causes Control 21: 77-82, 2010.
45. Engelhardt M, Ozkaynak MF, Drullinsky P, Sandoval C, Tugal O, Jayabose S and Moore MA: Telomerase activity and telomere length in pediatric patients with malignancies undergoing chemotherapy. Leukemia 12: 13-24, 1998.
46. Franco S, Ozkaynak MF, Sandoval C, Tugal O, Jayabose S, Engelhardt M and Moore MAS: Telomere dynamics in childhood leukemia and solid tumors: a follow-up study. Leukemia 17: 401-410, 2003.
47. Diker-Cohen T, Uziel O, Szyper-Kravitz M, Shapira H, Natur A and Lahav M: The effect of chemotherapy on telomere dynamics: clinical results and possible mechanisms. Leuk Lymphoma 54: 2023-2029, 2013.

Appendix 5.2 Table of telomere biology genes

Gene	Telomere Maintenance	Regulation of TERT	Extension of Telomeres	Mirabello *et al.*[318]
ABL1		●		
ACD	●	●		●
AKT1		●		
ATM		●		●
BICD1				●
BLM		●		●
CCND1		●		
CDKN1B		●		
CLPTM1L				●
DDX1				●
DDX11				●
DKC1	●	●	●	
DNA2	●		●	
E2F1		●		
E6		●		
EGF		●		
EGFR		●		
ESR1		●		
FEN1	●		●	
FOS		●		
H2AFB1	●			
H2AFX	●			
H2AFZ	●			
H2BFS	●			
HDAC1		●		
HDAC2		●		
HIST1H2AB	●			
HIST1H2AC	●			
HIST1H2AD	●			
HIST1H2AE	●			
HIST1H2AJ	●			
HIST1H2BA	●			
HIST1H2BB	●			
HIST1H2BC	●			
HIST1H2BD	●			
HIST1H2BE	●			
HIST1H2BF	●			
HIST1H2BG	●			
HIST1H2BH	●			
HIST1H2BI	●			
HIST1H2BJ	●			
HIST1H2BK	●			
HIST1H2BL	●			
HIST1H2BM	●			
HIST1H2BN	●			
HIST1H2BO	●			
HIST1H4A	●			
HIST1H4B	●			
HIST1H4C	●			
HIST1H4D	●			
HIST1H4E	●			
HIST1H4F	●			

Gene	Telomere Maintenance	Regulation of TERT	Extension of Telomeres	Mirabello *et al.*[318]
HIST1H4H	●			
HIST1H4I	●			
HIST1H4J	●			
HIST1H4K	●			
HIST1H4L	●			
HIST2H2AA3	●			
HIST2H2AA4	●			
HIST2H2AC	●			
HIST2H2BE	●			
HIST2H4A	●			
HIST2H4B	●			
HIST3H2BB	●			
HIST3H3	●			
HIST4H4	●			
HNRNPC		●		
HSP90AA1		●		
HUS1		●		
IFNAR2		●		
IFNG		●		
IL2		●		
IRF1		●		
JUN		●		
LIG1	●		●	
MAD1L1				●
MAPK1		●		
MAPK3		●		
MAX		●		
MEN1				●
MRE11A		●		●
MTOR		●		
MXD1		●		
MYC		●		●
NBN		●		●
NCL		●		
NFKB1		●		
NHP2	●		●	
NOLA1				●
NOLA2				●
NOLA3				●
NR2F2		●		
PARP1				●
PARP2		●		●
PCNA	●		●	
PIF1				●
PINX1		●		●
POLA1	●		●	
POLA2	●		●	
POLD1	●		●	
POLD2	●		●	
POLD3	●		●	
POLD4	●		●	
POLE	●		●	
POLE2	●		●	
POT1	●	●		●
PRIM1	●		●	
PRIM2	●		●	

Gene	Telomere Maintenance	Regulation of TERT	Extension of Telomeres	Mirabello *et al.*[318]
PRKDC				●
PTGES3		●		
RAD1		●		
RAD50		●		●
RAD51AP1				●
RAD51C				●
RAD51L1				●
RAD51L3				●
RAD54L				●
RAD9A		●		
RBBP4		●		
RBBP7		●		
RECQL				●
RECQL4				●
RECQL5				●
RFC1	●		●	
RFC2	●		●	
RFC3	●		●	
RFC4	●		●	
RFC5	●		●	
RPA1	●		●	
RPA2	●		●	
RPA3	●		●	
RPS6KB1		●		
RTEL1				●
RUVBL1	●		●	
RUVBL2	●		●	
SAP18		●		
SAP30		●		
SIN3A		●		
SIN3B		●		
SIP1				●
SMAD3		●		
SMG5		●		
SMG6		●		
SP1		●		
SP3		●		
TEP1				●
TERC	●	●	●	●
TERF1	●	●		●
TERF2	●	●		●
TERF2IP	●	●		●
TERT	●	●	●	●
TGFB1		●		
TINF2	●	●		●
TNKS		●		●
TNKS2				●
UBE3A		●		
WRAP53	●		●	
WRN		●		●
WT1		●		
XRCC5		●		
XRCC6		●		●
YWHAE		●		